WHEN AND WHY CAN TEST-TIME ADAPTATION SUCCEED?

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) modifies deployed models using only unlabeled target data, yet a fundamental theoretical question remains open: *when can unlabeled observations determine a better target predictor, and why can entropy-based updates correct errors rather than merely sharpen them?* We identify *recoverable semantic structure* as a common principle underlying both questions. First, we characterize conditions under which target semantics are recoverable from unlabeled data: semantic components must be recoverable and oriented by the source model, with within-semantic support dominating cross-semantic leakage. Under these conditions, we establish pointwise component-resolution learnability in the probably approximately correct (PAC) sense and derive finite-sample guarantees for exact component-level semantic identification. We then ask how entropy-based TTA algorithms exploit this information. For finite-width two-layer networks with a fixed output head, we show that entropy on an isolated prediction carries no correctness signal; correction instead arises from parameter sharing, which propagates source-oriented evidence through a parameter-gradient coupling kernel. Under a positive support-over-leakage condition controlling finite-width dynamics, we prove early-stage target-risk reduction. Guided by the same principle, we finally introduce **PR**ojected **I**nvariant **S**pectral **M**atching (PRISM), an optimization-free method that explicitly preserves shared task-relevant evidence while suppressing view-specific nuisance in the classifier-visible subspace. Across five distribution-shift benchmarks, PRISM consistently outperforms strong backpropagation-based and backpropagation-free TTA baselines. Code is available here.

1 INTRODUCTION

Deep models frequently suffer severe degradation under distribution shift, spurring rapid progress in test-time adaptation (TTA)—where a deployed model leverages unlabeled target data to adapt at inference time (Wang et al., 2020; Liang et al., 2025). Despite steady empirical advances spanning entropy minimization, selective sampling, and optimization-free heuristics (Niu et al., 2023; Murphy et al., 2026), the foundational reason why unlabeled adaptation succeeds at all remains poorly understood (Niu et al., 2026). While prior theoretical works analyze specific algorithms or idealized protocols (Farina et al., 2024; Gui et al., 2024; Zhou et al., 2026), they do not address the fundamental question of what structure the data must possess for test-time adaptation to be feasible at all. This limitation points directly to an information-theoretic barrier: as target labels are entirely absent, adaptation cannot succeed universally. Indeed, if two target distributions induce identical unlabeled observations but require conflicting semantic labels, no TTA algorithm can distinguish them (David et al., 2010). Therefore, test-time adaptation can succeed only if the target domain exhibits structure that is both identifiable from unlabeled data and correctly oriented by the pretrained source model. We formalize this prerequisite as *recoverable semantic structure*: target instances naturally form stable, near-pure components; the source model provides the necessary semantic orientation; and within-semantic support strictly dominates cross-semantic leakage.

Building on this recoverable semantic structure, we formulate pointwise component-resolution learnability in the probably approximately correct (PAC) sense for TTA. The key conditions are that semantic components can be recovered from unlabeled observations, their semantic orientation is supplied by the source model, and within-semantic support dominates cross-semantic leakage.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

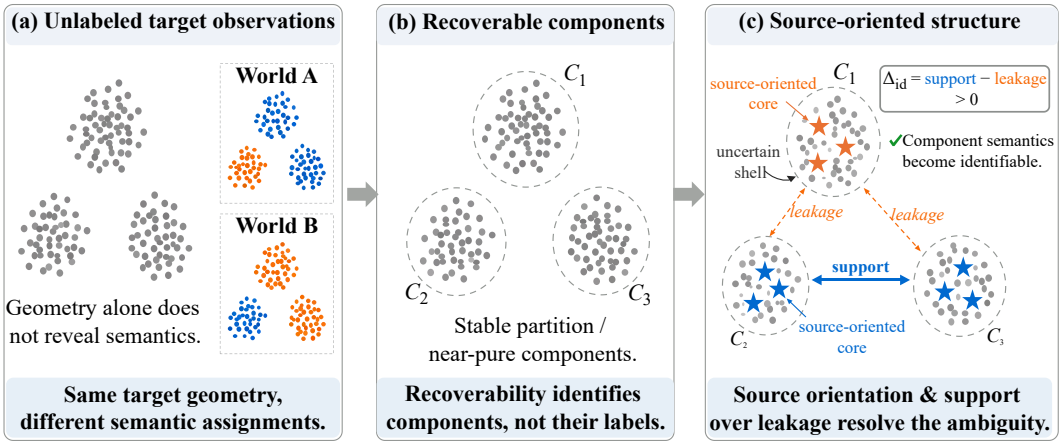

Figure 1: Learnability of unlabeled TTA. Unlabeled target data reveal components but not their semantic labels. Learnability arises when the source model orients these components and within-semantic support dominates cross-semantic leakage.

We prove that a positive effective identification margin Δ_{id} makes the target component labeling identifiable and recoverable from finitely many unlabeled samples.

Learnability alone, however, does not explain how existing TTA algorithms exploit this structure. We therefore analyze entropy minimization (Press et al., 2024; Wang et al., 2020). Entropy itself provides no correctness signal: for an isolated sample, it simply reinforces the current prediction. We show that error correction instead arises from *parameter sharing*, which propagates source-oriented evidence across samples through a parameter-gradient coupling kernel. For finite-width two-layer ReLU networks, we derive the exact update dynamics and prove early-stage risk reduction when within-semantic support dominates cross-semantic leakage, finite-width drift, and stochastic noise.

Together, the PAC and entropy analyses suggest a practical design principle: reliable task-relevant evidence should be preserved and aggregated while nuisance and semantic leakage are suppressed. Guided by this principle, we propose **PR**ojected **I**nvariant **S**pectral **M**atching (PRISM), an optimization-free realization in the classifier-visible subspace that explicitly aggregates shared evidence across deterministic target views. Our posterior calibration further provides a direct theoretical interface from the recoverable semantic structure to the output of PRISM, yielding a conditional risk bound through its realization gap; the complete result and fixed-graph audit are given in Appendix E.

Experiments audit the theory and evaluate the resulting method across CIFAR-10-C, CIFAR-100-C, and ImageNet-C (Hendrycks & Dietterich, 2019), ImageNet-R (Hendrycks et al., 2021), and ImageNet-Sketch (Wang et al., 2019). Concretely, the proposed structural diagnostics hold jointly in most audited CIFAR-C environments, while the effective identification margin closely tracks component-level semantic recoverability; finite-width experiments further support the proposed parameter-coupling mechanism. Across five benchmarks, PRISM achieves 69.36% average accuracy, compared with 63.33% for the strongest baseline.

In summary, our main contributions are: (1) **Identifiability:** A pointwise component-resolution PAC framework for unlabeled TTA, together with a posterior calibration that controls arbitrary probabilistic predictors. (2) **Propagation:** A theoretical analysis demonstrating that entropy-based error correction is driven by parameter-gradient coupling propagating source-oriented evidence across samples, rather than the entropy objective alone. (3) **Realization:** Guided by the same support-over-leakage principle, we develop PRISM, an optimization-free method that explicitly preserves shared task evidence and suppresses nuisance in the classifier-visible subspace; its emitted posterior is further connected to the PAC theory through the calibrated realization gap.

2 Related Work

Test-time adaptation. A major line of TTA minimizes predictive entropy, from Tent (Wang et al., 2020) to methods with filtering or stability mechanisms (Niu et al., 2023; Lee et al., 2024). Because

entropy favors confidence rather than correctness, it remains unclear why such updates can correct errors instead of reinforcing them. We identify parameter sharing as the source of corrective signal.

Theory of TTA. Most closely related, Zhou et al. (2026) studied recovery on non-stationary streams under an explicit proxy–task gradient-alignment condition. We instead ask when target semantics are identifiable from the source model and unlabeled observations, and show how entropy can acquire a corrective direction through parameter sharing rather than assumed gradient alignment.

Cluster assumptions and semantic structure. Our framework extends classical cluster and low-density assumptions (Yang et al., 2022): the semantic scale must be recoverable, source information must resolve label-permutation ambiguity, and within-semantic support must dominate cross-semantic leakage. Further discussion and positioning are deferred to Appendix F.

3 PAC Learnability of Test-Time Adaptation

We formalize the information-theoretic limit of test-time adaptation: when a frozen source model and unlabeled target observations suffice to identify target semantics, and when this recovery can improve upon the source predictor. We give sufficient conditions based on *recoverable semantic structure*: a near-pure recoverable partition, source-provided semantic orientation, and within-semantic support dominating cross-semantic leakage. Proofs, lower bounds, leave-one obstructions, and technical details are deferred to Appendix A.

3.1 Problem Setup and Assumptions

Source-Free TTA. Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = [K] := \{1, \dots, K\}$ the label space. The target distribution Q over $\mathcal{X} \times \mathcal{Y}$ induces an unlabeled marginal $\mu := Q_X$. For any hypothesis $h : \mathcal{X} \to [K]$, its target risk is defined as $L_Q(h) := Q(h(X) \neq Y)$. During deployment, the learner only accesses the frozen source model s, which provides predictions, confidence $\mathrm{conf}_s(x)$, representation $\varphi_s(x)$, and a symmetric affinity $W_s(x, x') \in [0, 1]$ measuring the similarity between two target inputs. An adaptation protocol Γ specifies which parameters may be updated at test time; letting $\Theta_\Gamma(s)$ be the reachable parameter space, the induced hypothesis class is $\mathcal{H}_s^\Gamma := \{h_{s,\theta} : \theta \in \Theta_\Gamma(s)\}$. A TTA learner is a mapping $A : (s, U_m) \mapsto \widehat{h} \in \mathcal{H}_s^\Gamma$ from unlabeled samples $U_m = (X_1, \dots, X_m) \sim \mu^m$.

Component-Resolution Learnability. Suppose $\mathcal{X} = C_0 \,\dot\cup\, \bigcup_{u \in \mathcal{V}} C_u$, where C_0 denotes an unresolved background region and $\mathcal{V}$ indexes the set of resolved semantic components. Let $m_u := \mu(C_u)$, $\zeta := \mu(C_0)$, and let $c(u) \in [K]$ denote the majority semantic label of C_u. The within-component label impurity is bounded by $Q(Y \neq c(u), X \in C_u) \leq \alpha_u m_u$, with aggregate impurity $\alpha_{\mathrm{tot}} := \sum_{u \in \mathcal{V}} \alpha_u m_u$. Let $\mathcal{H}_C$ denote the family of hypotheses that remain constant on every resolved component $\{C_u\}_{u \in \mathcal{V}}$. A discrete assignment $y = (y_u)_{u \in \mathcal{V}}$ is Γ-admissible if there exists $h \in \mathcal{H}_s^\Gamma \cap \mathcal{H}_C$ satisfying $h(x) = y_u$ for all $x \in C_u$. Letting $\mathcal{Y}_\Gamma(C)$ denote the set of admissible assignments and h_y its representative classifier (with fallback rules assigned on C_0), the component-resolution target is the resolved oracle risk: $L_{Q,C}^{\Gamma,\star} := \inf_{y \in \mathcal{Y}_\Gamma(C)} L_Q(h_y)$. An unconstrained oracle exploiting intra-component label variations is generally non-identifiable from purely marginal target data; Appendix A formalizes this via an information-theoretic two-world construction.

For a fixed source model s, we call TTA *pointwise component-resolution PAC learnable* with residual error $r(Q, s)$ if there exists a single learner A such that, for every target distribution Q satisfying the structural conditions below and protocol stability, and every $\varepsilon, \delta \in (0, 1)$, there exists a finite threshold $m_0(Q, s, \varepsilon, \delta)$ for which every $m \geq m_0(Q, s, \varepsilon, \delta)$ satisfies $\Pr_{U_m \sim Q_X^m}\left[L_Q(A(s, U_m)) \leq L_{Q,C}^{\Gamma,\star} + r(Q, s) + \varepsilon\right] \geq 1 - \delta$. The learner observes only (s, U_m); the unknown distribution Q enters the sample threshold through its structural parameters. We do not require a common sample threshold across all admissible target distributions. Our analysis establishes this guarantee with the irreducible residual error $r(Q, s) = \zeta + 2\alpha_{\mathrm{tot}}$.

Structural Assumptions. For a confidence threshold τ, let $A_j^\tau(s) := \{x \in \mathcal{X} : s(x) = j, \mathrm{conf}_s(x) \geq \tau\}$ denote the confident source anchor region for class j, and let $a_u^j := \mu(C_u \cap A_j^\tau(s))$ denote its target mass on C_u. The induced edge weights between components are given by $w_{uv} := \int_{C_u} \int_{C_v} W_s(x, x')\, d\mu(x) d\mu(x')$. For each class $k \in [K]$, define the true component index set $\mathcal{V}_k := \{u \in \mathcal{V} : c(u) = k\}$.

Assumption 3.1 (Recoverable Semantic Structure). **(A1) Near-Pure Partition.** $\mathcal{X} = C_0 \dot{\cup} \bigcup_{u\in\mathcal{V}} C_u$ with mass $\zeta = \mu(C_0)$, $m_u = \mu(C_u) > 0$, and within-component minority mass satisfying $Q(Y \neq c(u), X \in C_u) \leq \alpha_u m_u$. **(A2) Consistent Component Estimation.** There exists an unsupervised estimator $\mathcal{A}_0(s, D_1)$ yielding partitions $\{\widehat{C}_u\}$ such that the partition error $\delta_C := \max_{u\in\mathcal{V}} \mu(C_u \triangle \widehat{C}_{\pi(u)}) \xrightarrow{p} 0$ as $|D_1| \to \infty$ for some permutation π. **(A3) Source Orientation.** For every class $k \in [K]$ and every connected subset $S \subseteq \mathcal{V}_k$, $\partial_k(S) + \lambda A_k^{\mathrm{corr}}(S) \geq \kappa_{\mathrm{sup}} m(S)$, where $A_k^{\mathrm{corr}}(S) := \sum_{u\in S} a_u^k$, $m(S) := \sum_{u\in S} m_u$, and $\partial_k(S)$ denotes the boundary conductance within $\mathcal{V}_k$. **(A4) Support-Leakage Dominance.** Let the maximal normalized leakage rate be $\rho_{\mathrm{leak}} := \max_u \rho_u$, where $\rho_u := \frac{1}{m_u}\left(\sum_{v:c(v)\neq c(u)} w_{uv} + \lambda \max_{j\neq c(u)} a_u^j\right)$. The effective identification margin satisfies $\Delta_{\mathrm{id}} := \kappa_{\mathrm{sup}} - \rho_{\mathrm{leak}} > 0$.

A2 makes the semantic-resolution requirement explicit: the relevant component scale must be recoverable from unlabeled observations rather than chosen post hoc from semantic labels. Appendix A.9 gives a concrete label-free construction with a finite-sample recovery guarantee. A3–A4 govern the population graph–anchor functional $\mathcal{J}(y) := \sum_{u<v} w_{uv}\mathbf{1}\{y_u \neq y_v\} + \lambda \sum_{u\in\mathcal{V}} \sum_{j\neq y_u} a_u^j$. Let $c := (c(u))_{u\in\mathcal{V}}$ and $M(y) := \sum_u m_u \mathbf{1}\{y_u \neq c(u)\}$ denote the mass of misclassified components. Appendix A establishes $\mathcal{J}(y) - \mathcal{J}(c) \geq \Delta_{\mathrm{id}} M(y)$. Thus $\Delta_{\mathrm{id}} > 0$ makes the ground-truth component labeling the unique global minimizer of $\mathcal{J}$. Appendix A.7 further gives leave-one constructions showing that A1–A4 exclude distinct statistical obstructions, while incompatibility with the reachable protocol creates an irreducible risk floor, establishing the non-redundancy of these structural roles. The same margin also calibrates probabilistic outputs. For any predictor $p : \mathcal{X} \to \Delta_K$, let $\pi_u(p) := m_u^{-1} \int_{C_u} p(x)\, d\mu(x)$ and let $e(p)$ be its gap to the convex posterior relaxation of $\mathcal{J}$. Appendix E proves that, whenever $\Delta_{\mathrm{id}} > 0$, $L_Q(\arg\max p) \leq \zeta + \alpha_{\mathrm{tot}} + 2e(p)/\Delta_{\mathrm{id}}$. This provides a direct interface to PRISM without requiring it to explicitly optimize the PAC objective.

Protocol stability. To transfer component-level semantic identification under an estimated partition to a predictor reachable by the deployed protocol, we additionally assume that projection onto $\mathcal{Y}_\Gamma(C)$ is stable under partition perturbations of size δ_C. This is a protocol-level transfer condition, not a property of the target semantic structure.

3.2 Theoretical Guarantees

These structural conditions naturally induce a three-stage adaptation learner. Given an independent split of the unlabeled target sample into D_1 and D_2, the learner executes: (i) *Partition Recovery*, constructing the empirical component system $\widehat{C} = \mathcal{A}_0(s, D_1)$; (ii) *Component Identification*, obtaining discrete labelings via empirical risk minimization $\widehat{c} \in \arg\min_{y\in[K]^{|\widehat{\mathcal{V}}|}} \widehat{\mathcal{J}}(y)$ over D_2; and (iii) *Constrained Projection*, projecting $\widehat{c}$ onto the admissible hypothesis family under empirical mass weighting, $\widehat{y} \in \arg\min_{y\in\mathcal{Y}_\Gamma(\widehat{C})} \sum_u \widehat{m}_u \mathbf{1}\{y_u \neq \widehat{c}_u\}$, which yields the adapted predictor $\widehat{h} = h_{\widehat{y}}$.

Theorem 3.2 (Finite-Sample Generalization Bound). *Let $V := |\mathcal{V}|$ and $m_{\min} := \min_{u:m_u>0} m_u$. Suppose Assumption 3.1 and protocol stability hold, and the partition recovery satisfies $\delta_C \leq \min\{m_{\min}/2, \Delta_{\mathrm{id}}/(2C_{\mathrm{str}})\}$, where C_{str} is the structural perturbation constant defined in Appendix A. Then, with probability at least $1-\delta$ over the draw of $D_2 \sim \mu^m$,*

$$L_Q(\widehat{h}) \leq L_{Q,C}^{\Gamma,\star} + \zeta + 2\alpha_{\mathrm{tot}} + \mathcal{O}\left(V\sqrt{\frac{\log(V/\delta)}{m}} + \frac{(1+\lambda)^2\big(V\log K + \log(1/\delta)\big)}{(\Delta_{\mathrm{id}} - C_{\mathrm{str}}\delta_C)^2 m} + V\delta_C\right). \tag{1}$$

Furthermore, if $\delta_C = 0$, exact semantic identification ($M(\widehat{c}) = 0$) holds with probability at least $1-\delta$ provided that $m \gtrsim \frac{(1+\lambda)^2\left(V\log K + \log(1/\delta)\right)}{\Delta_{\mathrm{id}}^2 m_{\min}}$.

Theorem 3.2 shows that consistent component recovery ($\delta_C \to 0$) guarantees pointwise component-resolution PAC learnability. Conditioned on sufficiently accurate recovery, Stage 2 exact identification requires $\widetilde{O}(\Delta_{\mathrm{id}}^{-2} m_{\min}^{-1})$ samples, up to the displayed labeling-complexity and confidence factors. The Assouad lower bound in Appendix A.8 matches this leading dependence on Δ_{id} and $m_{\min}$, but not necessarily the V, K, logarithmic, or component-recovery factors. The full recover–identify–project

guarantee must additionally account for the Stage 1 samples required to drive the estimator-specific $\delta_C(m_1, \delta)$ below the structural transfer threshold; Appendix A.9 gives one explicit construction with a finite-sample recovery guarantee. The same bound also characterizes when identification improves upon the deployed source predictor. Let $\mathfrak{E}_m$ denote the finite-sample remainder in Theorem 3.2, and define $G_s(Q, C) := L_Q(s) - \left(L^{\Gamma,\star}_{Q,C} + \zeta + 2\alpha_{\text{tot}}\right)$. If $G_s(Q, C) > 0$ and $\mathfrak{E}_m < G_s(Q, C)$, then $L_Q(\widehat{h}) < L_Q(s)$. Since $\mathfrak{E}_m \to 0$ under consistent component recovery, every fixed problem with $G_s(Q, C) > 0$ eventually admits strict source improvement. Moreover, along any refining sequence of recoverable component systems with $L^{\Gamma,\star}_{Q,C} + \zeta + 2\alpha_{\text{tot}} \to 0$, the bound itself converges to zero.

3.3 Empirical Verification of Structural Conditions

We empirically audit whether the population structure required by our theory occurs on CIFAR-10-C and CIFAR-100-C (Hendrycks & Dietterich, 2019) across all 19 corruptions and 5 severity levels using a pretrained ResNet-50 (He et al., 2016).

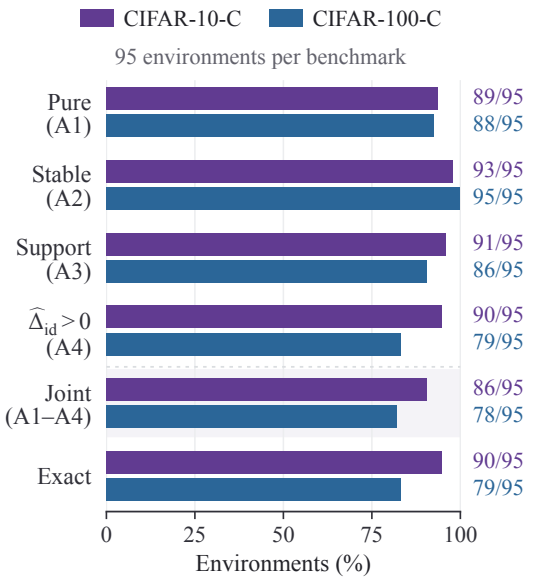

Figure 2: Prevalence of assumptions across 95 corruption–severity environments per benchmark.

Audit Protocol. In each environment, the 10,000 target instances are split into independent subsets D_1 (component estimation) and D_2 (graph and anchor construction). To test robustness to the partitioner, the benchmarks use different label-free component estimators: CIFAR-10-C uses source-prediction-constrained mutual-kNN components, whereas CIFAR-100-C uses a separate clustering construction from unlabeled AdaBN-recalibrated probabilities and features. Within each run, target labels are queried only *post-hoc* to audit purity, semantic support, leakage, and recovery. The reported dataset-level configurations were selected during exploratory development using label-based post-hoc audits, then used unchanged across all corruption–severity environments. We use common anchor parameters $\tau = 0.7$ and $\lambda = 0.05$; see Appendix B for dataset-specific constructions.

Empirical Prevalence and Identification Consistency. Figure 2 shows joint satisfaction of the four diagnostics in 90.5% of CIFAR-10-C and 82.1% of CIFAR-100-C environments. On the same frozen graph–anchor construction, all $169/190$ environments with $\widehat{\Delta}_{\text{id}} > 0$ recover the semantic component labeling, while all $21/190$ with $\widehat{\Delta}_{\text{id}} \leq 0$ fail. The positive direction agrees with the calibration theorem and confirms that the empirical optimizer reaches the unique semantic minimizer; the converse $21/21$ failure is empirical, not implied by the sufficient-condition theorem. Random, cyclic, and constant initializations converge to the same solution.

Finite-Sample Behavior. At severity 3, increasing m from 125 to 5000 raises exact recovery from 68.4% to 100% on CIFAR-10-C and from 0% to 84.2% on CIFAR-100-C. Across sizes and environments, $m[\widehat{\Delta}_{\text{id}}]^2_+$ correlates with exact recovery ($\rho = 0.884$) and error mass ($\rho = -0.844$), consistent with the predicted joint dependence on sample size and margin.

Orientation and Predictive Evidence. On CIFAR-100-C, over the same 95 environments and scoring samples, raw source, AdaBN-only, anchor-only, and full recover–identify–project attain 50.79%, 55.24%, 57.92%, and 61.72%. Thus the full pipeline gains $+6.48$ points beyond AdaBN and $+3.80$ beyond anchor-only. Among the 78 joint-diagnostic environments, 13 have $\widehat{G}_s > 0$, and all $13/13$ improve over the source; Appendix B gives details.

Holding the component graph fixed, $\lambda = 0$ yields 0% recovery over all severity-3 shifts, whereas $\lambda = 0.001$ restores $19/19$ CIFAR-10-C and $16/19$ CIFAR-100-C environments, supporting the role of source information in resolving the semantic permutation left by unlabeled geometry.

Finally, an observable label-free analogue of the effective margin uses only source outputs and unlabeled target features. Its configuration is selected from adaptation outcomes on severities $1, 3, 5$ and fixed for held-out severities $2, 4$; hence it is label-free at evaluation time but not a claim of universally unsupervised model selection. On the held-out severities, its Spearman correlation with adapted accuracy is 0.748–0.926 across Tent, SHOT, and AdaBN (Appendix B), providing external predictive evidence beyond the post-hoc structural audit.

4 Why Can Entropy Minimization Correct Errors?

The PAC analysis shows that unlabeled target data can contain recoverable semantics, but practical TTA methods do not explicitly solve the component-labeling problem. We further ask: *how can entropy minimization exploit this information?* For binary entropy, $\psi(s) := -H'(s) = s\sigma(s)(1-\sigma(s))$ has the same sign as s, so an isolated error is reinforced, not corrected. We show that correction arises from source-oriented evidence propagated across samples by shared parameters.

Problem Setup and Exact Finite-Width Coupling. We study the binary pure-component specialization of Section 3. Let $\mathcal{X}_T = \{x_1, \dots, x_N\}$ be the finite target pool with empirical measure μ and $\|x\|_2 \le 1$. Each semantic component C_u has orientation $\epsilon_u \in \{-1, +1\}$; $\epsilon(x) = \epsilon_u$ for $x \in C_u$ is used only to state and audit the theory, not by adaptation. For model score $s_t(x)$, define the signed margin $z_t(x) := \epsilon(x)s_t(x)$ and target risk $\mathcal{R}_t := \mu(\{x : z_t(x) \le 0\})$. Consider a width-$m_h$ two-layer ReLU network $s_t(x) = \sum_j a_j[w_{j,t}^\top x + b_{j,t}]_+$ with fixed output head $\{a_j\}$ and adapted hidden parameters ϑ_t. Its hidden-parameter coupling kernel is $K_t(x, x') := \langle \nabla_\vartheta s_t(x), \nabla_\vartheta s_t(x') \rangle = (1 + x^\top x') \sum_j a_j^2 g_{j,t}(x) g_{j,t}(x')$. Crucially, $\|x\|, \|x'\| \le 1$ gives $0 \le 1 + x^\top x' \le 2$; since a_j^2 and the gates are nonnegative, $K_t(x, x') \ge 0$ *pointwise* (See Appendix C). Define the influence field $F_t(x') := \int K_t(x, x')\psi(s_t(x))\, d\mu(x)$. Proposition C.2 proves the exact finite-width update: $s_{t+1}(x') - s_t(x') = \eta F_t(x') + \eta\xi_{t+1}(x') + R_{t+1}(x')$, where ξ_{t+1} is centered minibatch noise and R_{t+1} is the exact ReLU gate-switching remainder. Thus we can view K_t as an implicit sample graph: sender x influences receiver x' with strength $K_t(x, x')$.

Source-Oriented Support Implies a Corrective Field. For each component C_u, let $G_u \subseteq C_u$ be a *source-oriented core* satisfying $\epsilon_u s_0(x) > 0$, and let $B_u = C_u \setminus G_u$ be the uncertain shell. Lemma C.1 in Appendix C shows that a source-anchor margin plus an input-Lipschitz bound certifies such a core. Write $G_\epsilon := \bigcup_{u:\epsilon_u=\epsilon} G_u$ and $B := \bigcup_u B_u$. For a receiver component C_u, define its average signed initial field $\bar{F}_u := m_u^{-1} \int_{C_u} \epsilon_u F_0(x')\, d\mu(x')$. Because $\operatorname{sign}\psi(s_0(x)) = \operatorname{sign} s_0(x)$ and $K_0 \ge 0$, senders in G_{ϵ_u} contribute in the correct orientation, whereas opposite cores and shells can be treated adversarially. Define the corresponding support $S_u^J := m_u^{-1} \int_{C_u} \int_{G_{\epsilon_u}} K_0(x, x')|\psi(s_0(x))|\, d\mu(x) d\mu(x')$ and leakage $L_u^J := m_u^{-1} \int_{C_u} \int_{G_{-\epsilon_u} \cup B} K_0(x, x')|\psi(s_0(x))|\, d\mu(x) d\mu(x')$. Then $\bar{F}_u \ge S_u^J - L_u^J$. Let $\Delta_J := \min_u(S_u^J - L_u^J)$ and let $\omega_J := \max_u \sup_{x' \in C_u} |\epsilon_u F_0(x') - \bar{F}_u|$ measure within-component field variation. Proposition C.3 gives the component-to-pointwise bridge $d_0 := \min_{x' \in \mathcal{X}_T} \epsilon(x')F_0(x') \ge \Delta_J - \omega_J =: \underline{d}_0$. Hence component-level support yields a uniformly corrective initial field whenever $\Delta_J > \omega_J$, so that support dominates both leakage and within-component variation. The corrected margin $\underline{d}_0$ plays a dynamic role analogous to Δ_{id}.

Finite-width stability. A positive field at $t = 0$ is not sufficient by itself because ReLU gates and scores move during adaptation. For a movement budget Λ, define $A_2 := \sum_j a_j^2$, $\overline{\psi} := \sup_s |\psi(s)|$, and the initial gate-boundary energy $\Gamma_0(\Lambda) := \max_{x \in \mathcal{X}_T} \sum_j a_j^2 \mathbf{1}\{|u_{j,0}(x)| \le 2\Lambda|a_j|\overline{\psi}\}$. Here $\Gamma_0(\Lambda)$ measures mass close enough to a ReLU boundary to switch within the budget. Appendix C proves that $\eta(T+1) \le \Lambda$ implies $|K_t - K_0| \le 4\Gamma_0(\Lambda)$ and $|R_{t+1}| \le 2\eta\overline{\psi}\Gamma_0(\Lambda)$. With $\sigma_{\mathrm{drift}} := 2\Lambda A_2 \overline{\psi}$ and $\|\psi'\|_\infty \le 1/4$, define $\Delta_{\mathrm{corr}}(\Lambda) := \underline{d}_0 - 6\overline{\psi}\Gamma_0(\Lambda) - \frac{A_2}{2}\sigma_{\mathrm{drift}}$. Then we have:

Theorem 4.1 (Finite-Width Early Risk Descent). *If $\Delta_{\mathrm{corr}}(\Lambda) > 0$ and full-batch entropy gradient descent satisfies $\eta(T+1) \le \Lambda$, then every $x' \in \mathcal{X}_T$ obeys $z_T(x') \ge z_0(x') + \eta T \Delta_{\mathrm{corr}}(\Lambda)$. Thus no initially correct point becomes incorrect, every error with $-\eta T \Delta_{\mathrm{corr}}(\Lambda) < z_0(x') \le 0$ is corrected, and positive mass in this interval implies $\mathcal{R}_T < \mathcal{R}_0$.*

The minibatch extension adds a uniform martingale penalty; see Appendix C. Appendix C also provides one constructive sufficient route to a positive $\Delta_{\mathrm{corr}}(\Lambda)$ at finite width. We next complement this existence result by directly auditing whether the conditions entering the main theorem occur in naturally trained high-dimensional networks. In short, entropy supplies confidence amplification, source-oriented cores supply correctness, and parameter sharing supplies propagation.

4.1 Empirical Audit of the Corrective Regime

We instantiate the theory in a naturally trained finite-width setting. A frozen WideResNet (Zagoruyko & Komodakis, 2016) maps CIFAR-10/CIFAR-10-C images to 640-dimensional penultimate embeddings, classified by a width-512 two-layer ReLU network trained on clean source embeddings with ordinary stochastic gradient descent (SGD) and a fixed Rademacher output head. Source-only K-means defines components before target labels are inspected. Across two class pairs, two source checkpoints, and 42 corruption–severity environments, the frozen audit pool contains 1,018 semantically pure component unions with at least one initial error; see Appendix D.

Natural realization of corrective coupling. Among these 1,018 unions, 685 satisfy $\Delta_J > 0$, 262 satisfy $\Delta_J - \omega_J > 0$, and 392 have a uniformly positive exact initial field d_0. Under the common pre-specified movement budget $\Lambda = 10^{-3}$, 133/1,018 satisfy the complete pre-adaptation certificate $\Delta_{\mathrm{corr}}(\Lambda) > 0$ of Theorem 4.1. Thus naturally occurring unions realize the full sufficient regime, including the component-to-pointwise bridge and finite-width stability penalties.

The gap between the analytic certificate and exact field is expected: the former conservatively controls within-component field variation, moving ReLU gates, and score drift uniformly, whereas d_0 directly measures the realized semantic direction of the coupling graph.

Interventions on semantic support and leakage. Holding receivers fixed, we perturb the exact initial field to isolate the mechanism. Among the 392 cases with $d_0 > 0$, core-only senders preserve positivity in 392/392, whereas removing the cores, using only shells, or shuffling semantic orientations yields 0/392 positive fields. Multiplying only cross-semantic coupling by $1.25\times$, $1.5\times$, $2\times$, and $4\times$ reduces the positive-field fraction monotonically to 72.70%, 49.49%, 22.70%, and 1.28%. These interventions directly support the predicted support-over-leakage mechanism in the realized coupling graph.

The PAC analysis establishes when semantic information is statistically identifiable; the finite-width analysis shows how parameter sharing propagates source-oriented evidence into corrective score motion. This motivates replacing implicit iterative propagation with an explicit mechanism that preserves shared task evidence while suppressing nuisance, which we develop next.

5 Projected Invariant Spectral Matching

The preceding analysis motivates preserving task-relevant evidence while suppressing nuisance rather than relying on iterative parameter updates to discover this balance implicitly. Following this principle, we introduce **PR**ojected **I**nvariant **S**pectral **M**atching (PRISM), an optimization-free TTA method that treats deterministic target views as repeated measurements in the classifier-visible space. This theory-guided design connection is complemented by the posterior calibration in Appendix E, which provides a direct mathematical interface from recoverable semantic structure to the actual output of PRISM. Figure 3 summarizes the pipeline.

5.1 Methodology

Classifier-visible repeated measurements. Let the frozen network produce feature $h \in \mathbb{R}^D$ and logits $s = Wh + b$. Since only relative logit margins affect predictions, we therefore center the classifier with $W_c = W - K^{-1}\mathbf{1}\mathbf{1}^\top W$, and let $W_c = U\Sigma B^\top$ be its compact singular value decomposition (SVD), where $r = \mathrm{rank}(W_c) \le K - 1$ and $A := W_c B$. For sample i under view v, define $\widetilde{s}_{iv} = (s_{iv} - b) - K^{-1}\mathbf{1}\mathbf{1}^\top(s_{iv} - b)$, $z_{iv} = \widetilde{s}_{iv}(A^\top)^\dagger \in \mathbb{R}^r$. Indeed, any displacement in $\ker(W_c)$ changes logits only by a class-shared constant and hence leaves softmax unchanged, thus all prediction-relevant correction can be represented by z_{iv}. Our standard ViT-B five-benchmark evaluation uses the same $V = 5$ deterministic views on every dataset. Cross-backbone transfer controls use the view budgets explicitly reported with each architecture; in particular, the ViT-S and

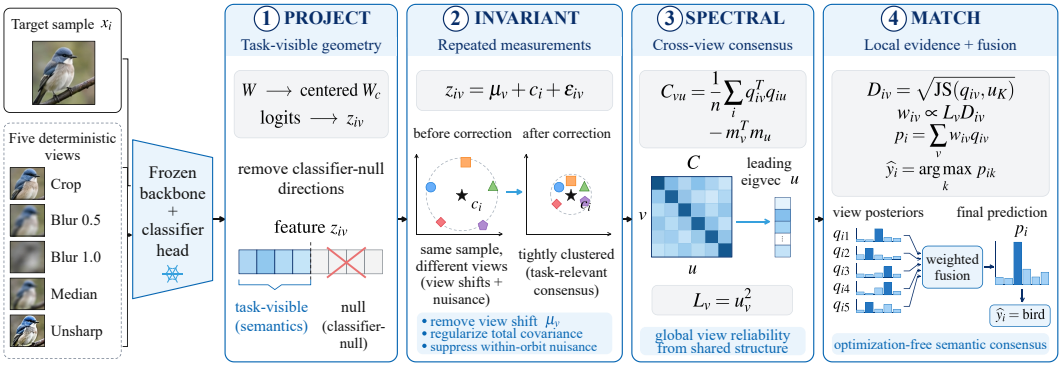

Figure 3: Overview of PRISM. PRISM removes view-specific offsets, corrects target anisotropy, and suppresses within-orbit nuisance.

ViT-L/16 transfer evaluations use four-view PRISM variants. The views are treated as repeated measurements $z_{iv} = \mu_v + c_i + \varepsilon_{iv}$, where μ_v is a systematic view shift, c_i is shared task-relevant variation, and ε_{iv} is view-specific nuisance.

Causal invariant transport. For the currently observed target prefix of n samples, let $\mu_v = \frac{1}{n}\sum_i z_{iv}, \widehat{z}_{iv} = z_{iv} - \mu_v, S_T = \frac{1}{nV}\sum_{i,v} \widehat{z}_{iv}\widehat{z}_{iv}^\top$. We stabilize the target covariance by Oracle Approximating Shrinkage (OAS) (Chen et al., 2010). Writing $\bar{\lambda} = \mathrm{Tr}(S_T)/r$ and $\alpha = \|S_T\|_F^2 / r^2$, we use $\rho = \min\left\{1, \frac{\alpha+\bar{\lambda}^2}{(n+1)(\alpha-\bar{\lambda}^2/r)}\right\}, \Sigma_T = (1-\rho)S_T + \rho\bar{\lambda}I$. Let $\Sigma_T = U_T \,\mathrm{diag}(\lambda_1, \ldots, \lambda_r) U_T^\top$ and $\lambda_{\mathrm{avg}} = r^{-1}\sum_j \lambda_j$. With $f_j = \sqrt{\lambda_{\mathrm{avg}}/\lambda_j}$ and $\kappa_n = n/(n+r)$, the finite-sample total-covariance transform is $a_j = f_j^{\kappa_n}$ if $\lambda_j < \lambda_{\mathrm{avg}}$, otherwise $f_j; T_n = U_T \,\mathrm{diag}(a_1, \ldots, a_r) U_T^\top$. To separate shared variation from augmentation-induced nuisance, define $\bar{z}_i = V^{-1}\sum_v z_{iv}$ and $\mu = V^{-1}\sum_v \mu_v$. The across-sample and within-orbit covariances are $S_B = \frac{1}{n}\sum_i \bar{z}_i\bar{z}_i^\top - \mu\mu^\top, S_W = S_T - S_B$. For $S_W = U_W \,\mathrm{diag}(\eta_1, \ldots, \eta_r) U_W^\top$ and $\eta_{\mathrm{avg}} = r^{-1}\sum_j \eta_j$, we suppress only unusually large within-orbit variation: $c_j = \min\left\{\sqrt{\eta_{\mathrm{avg}}/\eta_j}, 1\right\}, N_n = U_W \,\mathrm{diag}(c_1, \ldots, c_r) U_W^\top$. The final transport and corrected posterior are $G_n = N_n^{1/2} T_n N_n^{1/2}, q_{iv} = \mathrm{softmax}\big((z_{iv} - \mu_v) G_n A^\top + b\big)$. Here q_{iv} denotes the posterior emitted when sample i arrives, using the current-batch-inclusive prefix available at that time; each emitted posterior remains fixed after later batches update μ_v or G_n. The fully time-indexed recursion is given in Appendix G.8. Thus T_n corrects global target anisotropy while N_n selectively contracts directions dominated by within-orbit nuisance.

Centered spectral semantic consensus. A reliable view should agree with the others on a sample-dependent signal, rather than merely share a biased class marginal. Let $m_v = n^{-1}\sum_i q_{iv}$ and define the centered posterior collision matrix $C_{vu} = \frac{1}{n}\sum_i q_{iv}^\top q_{iu} - m_v^\top m_u$. Let u be its unit leading eigenvector. We define the global view leverage $L_v = u_v^2$ and the local evidence $D_{iv} = \sqrt{\mathrm{JS}(q_{iv}, u_K)}$ relative to the uniform posterior u_K. Prediction is obtained by $w_{iv} = \frac{L_v D_{iv}}{\sum_j L_j D_{ij}}, p_i = \sum_v w_{iv} q_{iv}, \widehat{y}_i = \arg\max_k p_{ik}$. Under the repeated-measurement model $q_{iv} - m_v = a_v g_i + \xi_{iv}$ with centered cross-view-independent noise, the population collision matrix contains a rank-one shared component proportional to $aa^\top$ plus view-specific diagonal noise. Hence, when this shared component has a leading eigengap, the principal eigenvector estimates its loading direction up to sign. We use this result only to motivate L_v; the full statement is given in Appendix G.14.

5.2 Evaluation

Protocol and baselines. We follow Murphy et al. (2026) and additionally evaluate CIFAR-100-C under the corresponding setting. Table 1 compares PRISM with T3A (Iwasawa & Matsuo, 2021), SAR (Niu et al., 2023), LAME (Boudiaf et al., 2022), Tent (Wang et al., 2020), CoTTA (Wang et al., 2022), FOA (Niu et al., 2024), Surgeon (Ma et al., 2025), and NEO (Murphy et al., 2026). For the four shared benchmarks, we use the published baselines of Murphy et al. (2026); all baselines are also evaluated under our standardized implementation, with matched-sample, long-stream, and matched-view NEO controls in Appendix G.10.

Table 1: Accuracy (%) on the ViT-B benchmark setting. † denotes published baseline values from Murphy et al. (2026) on the four shared benchmarks. PRISM reports mean±SD over three evaluation seeds. Matched-implementation and equal-view controls are in Appendix G.10.

Dataset	No Adapt	T3A	SAR	LAME	Tent	CoTTA	FOA	Surgeon	NEO	PRISM
ImageNet-C†	55.60	55.20	55.90	54.20	56.30	55.90	58.40	56.10	59.20	**62.77** ±0.48
CIFAR-10-C†	80.40	80.10	80.60	79.80	81.30	80.60	80.90	82.70	82.40	**89.87** ±0.13
ImageNet-R†	59.20	58.70	59.30	58.50	59.40	59.20	60.20	60.20	60.30	**65.69** ±0.18
ImageNet-Sketch†	45.40	45.50	45.50	45.00	45.70	45.20	46.30	47.00	47.20	**53.26** ±0.79
CIFAR-100-C	61.56	63.03	61.56	55.51	61.39	61.60	66.80	58.88	67.53	**75.22** ±0.62
Avg	60.43	60.51	60.57	58.60	60.82	60.50	62.52	60.98	63.33	**69.36**

Table 2: **Ablation, architecture transfer, and posterior bridge of PRISM.** (a) Stepwise decomposition on the five benchmarks; all variants use the same five-view stream, and the complete PRISM row corresponds to the standard 512-sample configuration in Table 1. (b) ImageNet-C architecture transfer. The standard ViT-B and ResNet-50 PRISM evaluations use five views, whereas the ViT-S and ViT-L/16 transfer evaluations use four views; NEO values are the published architecture references from Murphy et al. (2026). (c) Fixed-graph posterior bridge on the 512-sample audit; accuracy is the 47-domain audit-split average, not the five-benchmark macro average.

(a) Component ablation

Variant	C10-C	C100-C	IN-C	IN-R	IN-S	Macro
Frozen source	79.21	61.56	54.82	58.01	44.27	59.57
Raw five-view average	81.25	66.41	56.05	60.16	48.04	62.38
+ Classifier-visible re-expression	81.25	66.41	56.05	60.16	48.04	62.38
+ Total-covariance transport	86.18	71.46	60.66	62.70	50.39	66.28
+ View-wise centering	88.32	73.22	61.65	64.65	50.98	67.76
+ Within-orbit suppression	89.54	74.39	62.73	64.45	51.95	68.61
+ Centered spectral leverage	89.02	74.96	62.34	65.43	52.54	68.86
+ Local JS evidence (PRISM)	**89.87**	**75.22**	**62.77**	**65.69**	**53.26**	**69.36**

(b) Backbone transfer on ImageNet-C

Backbone	PRISM views	NEO	PRISM	Gain
ResNet-50	5	25.66	**38.12**	+12.46
ViT-S	4	43.20	**45.47**	+2.27
ViT-B	5	59.20	**62.77**	+3.57
ViT-L	4	63.10	**64.74**	+1.64

(c) Fixed-graph posterior bridge

Method	Accuracy ↑	$e(p)$ ↓	M_{soft} ↓
Source	66.17	5.69×10^{-3}	0.1835
PRISM	**75.07**	2.03×10^{-3}	**0.1056**

Ablation and architecture transfer. Table 2(a) raises the five-benchmark macro from 62.38% for raw five-view averaging to 66.28%, 67.76%, 68.61%, 68.86%, and finally 69.36% after total-covariance transport, view-wise centering, nuisance suppression, spectral leverage, and local JS evidence, respectively, for a total +6.98 points. The same formulation transfers across architectures under the reported view budgets (Table 2(b)), exceeding published NEO on ImageNet-C by +12.46, +2.27, +3.57, and +1.64 points for ResNet-50 and ViT-S/B/L, respectively.

PAC–PRISM output bridge. On the same frozen component graph and source anchors, the 512-sample audit shows that PRISM improves 47-domain accuracy by 8.90 points while reducing the mean realization gap from 5.69×10^{-3} to 2.03×10^{-3}, with a lower gap in 42/47 domains; the expanded 4,096-sample audit gives the same conclusion in 45/47. Theory, protocol, uncertainty intervals, and raw-five-view controls are in Appendix E.

Efficiency. PRISM requires no backward gradient. Under matched protocol, it exceeds NEO by +6.33 and +4.98 points on the 512- and 4,096-sample settings, and under strict equal-view control still exceeds NEO-5V by +5.40 and +4.78 points. Its persistent state is $O(r^2 + Vr + VK + V^2)$ and the dominant classifier-visible spectral update costs $O(r^3)$. Matched-protocol/view controls and runtime/memory measurements are reported in Appendices G.10 and G.17, respectively.

6 Conclusion

We identified recoverable semantic structure as a unifying principle for unlabeled test-time adaptation. We first established a pointwise component-resolution PAC framework showing when source-oriented semantic structure could be identified from unlabeled target observations. We then showed that entropy minimization itself provided no correctness signal and that corrective dynamics instead arose from source-oriented evidence propagated through parameter sharing. Guided by the same support-over-leakage principle, we developed PRISM, an optimization-free method that preserved shared task-relevant evidence while suppressing nuisance in the classifier-visible subspace. Together, these results linked statistical identifiability, finite-width correction dynamics, and practical adaptation within a common semantic-structure perspective.

AI USE STATEMENT

Large language models were used to assist with language polishing and to check theoretical derivations and code. The core research ideas, original mathematical derivations, and original code implementation were developed by the authors. All AI-assisted suggestions were independently reviewed and verified by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We support reproducibility in several ways. First, all assumptions, proofs, lower bounds, and auxiliary results for the PAC and finite-width analyses are provided in Appendices A and C, with the posterior calibration results in Appendix E. Second, the complete specification of PRISM, including its causal update rules, deterministic views, implementation details, hyperparameters, evaluation protocols, baseline provenance, and computational measurements, is provided in Appendix G; additional experiments auditing the theoretical conditions are reported in Appendices B and D. Finally, anonymized source code for reproducing the experiments is provided at https://anonymous.4open.science/r/tta_theory-1ADB/readme.

REFERENCES

Malik Boudiaf, Romain Müller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8334–8343. IEEE, 2022.

Tianle Cai, Ruiqi Gao, Jason Lee, and Qi Lei. A theory of label propagation for subpopulation shift. In *International Conference on Machine Learning*, pp. 1170–1182. PMLR, 2021.

Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. *Advances in Neural Information Processing Systems*, 33:21061–21071, 2020.

Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 129–136. JMLR Workshop and Conference Proceedings, 2010.

Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093, 2024.

Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and J Zico Kolter. Test time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 35:6204–6218, 2022.

Shurui Gui, Xiner Li, and Shuiwang Ji. Active test-time adaptation: Theoretical analyses and an algorithm. *arXiv preprint arXiv:2404.05094*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 8320–8329. IEEE, 2021.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440, 2021.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pp. 5468–5479. PMLR, 2020.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, volume 2024, pp. 46977–47002, 2024.

Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6028–6039. PMLR, 2020.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Ke Ma, Jiaqi Tang, Bin Guo, Fan Dang, Sicong Liu, Zhui Zhu, Lei Wu, Cheng Fang, Ying-Cong Chen, Zhiwen Yu, and Yunhao Liu. SURGEON: memory-adaptive fully test-time adaptation via dynamic activation sparsity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 30514–30523. Computer Vision Foundation / IEEE, 2025.

Alexander Murphy, Michal Danilowski, Soumyajit Chatterjee, and Abhirup Ghosh. Neo—no-optimization test-time adaptation through latent re-centering. In *International Conference on Learning Representations*, volume 2026, pp. 128762–128789, 2026.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. *arXiv preprint arXiv:2404.01650*, 2024.

Shuaicheng Niu, Guohao Chen, Yaofo Chen, Zhiquan Wen, Jinwu Hu, Zeshuai Deng, Deyu Chen, Shuhai Zhang, Renjie Chen, Zihao Lian, et al. A survey on self-improving test-time intelligence: Feedback-driven adapting, learning, and scaling at inference. *arXiv preprint arXiv:2609.01679*, 2026.

Ori Press, Ravid Shwartz-Ziv, Yann LeCun, and Matthias Bethge. The entropy enigma: Success and failure of entropy minimization. *arXiv preprint arXiv:2405.05012*, 2024.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pp. 443–450. Springer, 2016.

Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. In *Domain adaptation in computer vision applications*, pp. 153–171. Springer, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 7191–7201. IEEE, 2022.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2022.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhi Zhou, Ming Yang, Shi-Yu Tian, Kun-Yang Yu, Lan-Zhe Guo, and Yu-Feng Li. On the learnability of test-time adaptation: A recovery complexity perspective. *arXiv preprint arXiv:2605.28057*, 2026.

A Proofs for PAC Learnability under Recoverable Semantic Structure

A.1 Preliminaries and proof structure

This appendix proves Theorem 3.2 and gives the accompanying necessity and lower-bound results. We use the notation of Section 3, but restate the complete statistical model and the additional objects used only in the proofs.

Source-free TTA and the PAC target. Let $\mathcal{X}$ be the input space and $\mathcal{Y} = [K]$ the label space. The target joint distribution is Q on $\mathcal{X} \times \mathcal{Y}$, with unlabeled marginal $\mu := Q_X$. For $h : \mathcal{X} \to [K]$, write $L_Q(h) := Q(h(X) \neq Y)$. At test time, the learner has access to a fixed source model s but neither the source training data nor target labels. The source model supplies its prediction $s(x)$, confidence $\mathrm{conf}_s(x)$, representation $\varphi_s(x)$, and a bounded symmetric affinity $W_s(x, x') \in [0, 1]$. An adaptation protocol Γ specifies the parameter states reachable from s; its induced hypothesis class is $\mathcal{H}_s^\Gamma := \{h_{s,\theta} : \theta \in \Theta_\Gamma(s)\}$. A TTA learner is a measurable map $A : (s, U_m) \mapsto \widehat{h} \in \mathcal{H}_s^\Gamma$, where $U_m = (X_1, \ldots, X_m) \sim \mu^m$ is unlabeled.

For a fixed source model s, we call TTA *pointwise component-resolution PAC learnable* with residual error $r(Q, s)$ if there exists a single learner A such that, for every target distribution Q satisfying the structural conditions below and protocol stability, and every $\varepsilon, \delta \in (0, 1)$, there exists a finite threshold $m_0(Q, s, \varepsilon, \delta)$ such that, for every $m \geq m_0(Q, s, \varepsilon, \delta)$,

$$\Pr_{U_m \sim Q_X^m}\Big[L_Q(A(s, U_m)) \leq L_{Q,C}^{\Gamma,\star} + r(Q, s) + \varepsilon\Big] \geq 1 - \delta. \tag{2}$$

The same learner is used for every admissible Q and observes only (s, U_m). The threshold may depend on the unknown distribution through its structural parameters; it is not supplied to the learner. No common sample threshold over all admissible distributions is required. Theorem 3.2 establishes this pointwise guarantee with $r(Q, s) = \zeta + 2\alpha_{\mathrm{tot}}$.

Semantic components and protocol-admissible decisions. Assumption 3.1 posits a decomposition $\mathcal{X} = C_0 \,\dot\cup\, \bigcup_{u \in \mathcal{V}} C_u$. The unresolved region has mass $\zeta := \mu(C_0)$; each resolved component has mass $m_u := \mu(C_u) > 0$ and dominant label $c(u) \in [K]$, with $Q(Y \neq c(u), X \in C_u) \leq \alpha_u m_u$. We write $\alpha_{\mathrm{tot}} := \sum_u \alpha_u m_u$, $V := |\mathcal{V}|$, and $m_{\min} := \min_u m_u$.

Let $\mathcal{H}_C$ be the class of hypotheses that are constant on every resolved component. A labeling $y = (y_u)_{u \in \mathcal{V}} \in [K]^{\mathcal{V}}$ is Γ-admissible if some $h \in \mathcal{H}_s^\Gamma \cap \mathcal{H}_C$ satisfies $h(x) = y_u$ on each C_u. We denote the set of such labelings by $\mathcal{Y}_\Gamma(C)$ and fix a representative h_y for each $y \in \mathcal{Y}_\Gamma(C)$; on C_0, h_y follows the protocol-specific fallback rule. The component-resolution oracle is

$$L_{Q,C}^{\Gamma,\star} := \inf_{y \in \mathcal{Y}_\Gamma(C)} L_Q(h_y). \tag{3}$$

Let $c := (c(u))_{u\in\mathcal{V}}$. For later use, define the semantic mismatch mass and the best reachable mismatch by

$$M(y) := \sum_{u\in\mathcal{V}} m_u \mathbf{1}\{y_u \neq c(u)\}, \qquad \beta_\Gamma(C) := \inf_{y\in\mathcal{Y}_\Gamma(C)} M(y). \tag{4}$$

The restriction to component-constant decisions is necessary for a uniform unlabeled guarantee. Theorem A.19 shows that an unrestricted reachable oracle can exploit label variation inside a component that is not identifiable from unlabeled target observations.

Source anchors and the component graph. Fix a confidence threshold τ and let $A_j^\tau(s) := \{x : s(x) = j,\ \mathrm{conf}_s(x) \geq \tau\}$ be the confident source-anchor region for class j, with component mass $a_u^j := \mu(C_u \cap A_j^\tau(s))$. Because $s(x)$ denotes a single predicted class, the anchor regions $\{A_j^\tau(s)\}_{j=1}^K$ are pairwise disjoint. Let $\lambda \geq 0$ denote the anchor coefficient. The source affinity induces

$$w_{uv} := \int_{C_u}\int_{C_v} W_s(x,x')\, d\mu(x) d\mu(x'). \tag{5}$$

For $k \in [K]$, let $\mathcal{V}_k := \{u : c(u) = k\}$. For $S \subseteq \mathcal{V}_k$, define

$$m(S) := \sum_{u\in S} m_u, \qquad \partial_k(S) := \sum_{u\in S}\sum_{v\in\mathcal{V}_k\setminus S} w_{uv}, \qquad A_k^{\mathrm{corr}}(S) := \sum_{u\in S} a_u^k. \tag{6}$$

The within-class support graph on $\mathcal{V}_k$ contains an edge whenever $w_{uv} > 0$. For each component, define

$$\rho_u := \frac{1}{m_u}\left(\sum_{v:c(v)\neq c(u)} w_{uv} + \lambda \max_{j\neq c(u)} a_u^j\right), \qquad \rho_{\mathrm{leak}} := \max_u \rho_u, \tag{7}$$

and recall $\Delta_{\mathrm{id}} := \kappa_{\mathrm{sup}} - \rho_{\mathrm{leak}}$. The population graph–anchor objective is

$$\mathcal{J}(y) = \sum_{u<v} w_{uv}\mathbf{1}\{y_u \neq y_v\} + \lambda \sum_{u\in\mathcal{V}}\sum_{j\neq y_u} a_u^j. \tag{8}$$

Recoverable semantic structure. For completeness, we restate the four structural ingredients of Assumption 3.1 in the form used in the proofs.

- **(A1)** *Near-pure partition.* The component system satisfies $Q(Y \neq c(u), X \in C_u) \leq \alpha_u m_u$ and has unresolved mass ζ.
- **(A2)** *Consistent component estimation.* From an unlabeled split D_1, an estimator $\mathcal{A}_0(s, D_1)$ returns $\widehat{C} = \{\widehat{C}_{\hat{u}}\}_{\hat{u}\in\widehat{\mathcal{V}}}$ such that, with probability tending to one, there is a bijection $\pi : \mathcal{V} \to \widehat{\mathcal{V}}$ for which $\delta_C := \max_u \mu(C_u \triangle \widehat{C}_{\pi(u)}) \to 0$.
- **(A3)** *Source orientation.* For every class k and every nonempty connected $S \subseteq \mathcal{V}_k$, $\partial_k(S) + \lambda A_k^{\mathrm{corr}}(S) \geq \kappa_{\mathrm{sup}} m(S)$ for some $\kappa_{\mathrm{sup}} > 0$.
- **(A4)** *Support–leakage dominance.* $\Delta_{\mathrm{id}} = \kappa_{\mathrm{sup}} - \rho_{\mathrm{leak}} > 0$.

Protocol-transfer condition. In addition to A1–A4, the complete risk guarantee uses protocol stability to transfer a statistically identified component labeling from the population component system C to the estimated system $\widehat{C}$ while preserving protocol admissibility. This condition concerns the deployment protocol rather than the semantic structure of the target distribution and is therefore kept separate from Assumption 3.1.

Recover–identify–project learner. The proof analyzes the three-stage learner from Section 3.2. The unlabeled sample is split into independent sets D_1 and D_2. First, $\widehat{C} = \mathcal{A}_0(s, D_1)$ is recovered from D_1. Second, using D_2, the learner minimizes the empirical analogue of (8) over the full component-labeling space to obtain an identified labeling $\widehat{c}$. Finally, $\widehat{c}$ is projected onto $\mathcal{Y}_\Gamma(\widehat{C})$ using empirical component masses, yielding a reachable predictor $\widehat{h}$.

Proof structure. The argument is separated according to the role of the assumptions. Section A.2 converts component mismatch into target risk and isolates the resolution error from A1. Section A.3 proves the population calibration inequality $\mathcal{J}(y) - \mathcal{J}(c) \geq \Delta_{\mathrm{id}} M(y)$ from A3–A4. Section A.4 establishes finite-sample identification for a fixed component system. Section A.5 projects the identified labels onto the Γ-reachable class, and Section A.6 uses A2 and protocol stability to transfer the result to an estimated component system. Section A.7 gives leave-one counterexamples for A1–A4 and protocol stability, Section A.8 gives the matching dependence on $\Delta_{\mathrm{id}}^{-2} m_{\min}^{-1}$, and Section A.9 records concrete sufficient certificates for A2–A4.

A.2 Component mismatch and target risk

We first relate component-level mismatch to the target risk. This separates the statistical identification problem from the irreducible error associated with the chosen semantic resolution: even exact recovery of the dominant component labels cannot remove the unresolved mass in C_0 or the minority labels within a near-pure component.

Lemma A.1 (Risk and component mismatch)**.** *For every reachable component labeling y,*

$$M(y) - \alpha_{\mathrm{tot}} \leq L_Q(h_y) \leq \zeta + \alpha_{\mathrm{tot}} + M(y). \tag{9}$$

Consequently, $\beta_\Gamma(C) \leq L_{Q,C}^{\Gamma,\star} + \alpha_{\mathrm{tot}}$.

Proof. On C_0 the risk is at most ζ. On a resolved component C_u, if $y_u = c(u)$ then the error mass is at most $\alpha_u m_u$; if $y_u \neq c(u)$, the trivial upper bound is m_u. Summing gives the upper bound. Conversely, when $y_u \neq c(u)$, at least $(1 - \alpha_u) m_u$ mass in C_u has label $c(u)$ and is misclassified, giving the lower bound. Rearranging the lower bound yields $M(y) \leq L_Q(h_y) + \alpha_{\mathrm{tot}}$; taking the infimum over reachable y proves the final statement. □

The lemma isolates the resolution term that appears in the main theorem. The contribution $\zeta + 2\alpha_{\mathrm{tot}}$ is entirely due to unresolved mass and within-component minority labels; it is independent of the concentration and projection arguments below.

A.3 Population identifiability

We next establish the population calibration inequality used throughout the finite-sample analysis. Relative to the semantic labeling c, changing a set of component labels incurs within-class boundary and correct-anchor costs. Only cross-semantic edges that become uncut and wrong-oriented anchors can offset this increase. Assumptions A3–A4 ensure that the net change remains positive at a rate controlled by Δ_{id}. For a candidate y, define the cross-semantic edge gain and wrong-anchor gain

$$\Lambda_{\mathrm{cross}}(y) := \sum_{u<v:\, c(u)\neq c(v)} w_{uv} \mathbf{1}\{y_u = y_v\}, \tag{10}$$

$$A_{\mathrm{wrong}}(y) := \sum_{u:\, y_u \neq c(u)} a_u^{y_u}, \tag{11}$$

and $B_-(y) := \Lambda_{\mathrm{cross}}(y) + \lambda A_{\mathrm{wrong}}(y)$.

Lemma A.2 (Self-bounded leakage)**.** *With ρ_{leak} from Assumption 3.1,*

$$B_-(y) \leq \rho_{\mathrm{leak}} M(y) \qquad \forall y \in [K]^{\mathcal{V}}. \tag{12}$$

Proof. If a cross-semantic edge $\{u, v\}$ becomes uncut under y, at least one endpoint is mismatched; otherwise $y_u = c(u)$ and $y_v = c(v)$ would imply $c(u) = c(v)$. Charge each such edge to one mismatched endpoint. Then

$$\Lambda_{\mathrm{cross}}(y) \leq \sum_{u:y_u\neq c(u)} \sum_{v:c(v)\neq c(u)} w_{uv}.$$

Moreover $A_{\mathrm{wrong}}(y) \leq \sum_{u:y_u\neq c(u)} \max_{j\neq c(u)} a_u^j$. The definition of ρ_{leak} now gives $B_-(y) \leq \rho_{\mathrm{leak}} \sum_{u:y_u\neq c(u)} m_u$. □

Lemma A.3 (Connected-set reduction). *For a fixed class k, the minimum of $[\partial_k(S) + \lambda A_k^{\mathrm{corr}}(S)]/m(S)$ over all nonempty $S \subseteq \mathcal{V}_k$ equals the minimum over connected subsets of the within-class positive-weight graph.*

Proof. Decompose any nonempty S into connected components $S_1, \dots, S_r$ of the induced positive-weight graph. There are no positive-weight within-class edges between different S_i, so the boundary, anchor mass, and component mass are all additive across i. The ratio for S is therefore the $m(S_i)$-weighted average of the ratios for the S_i and cannot be smaller than their minimum. □

Proposition A.4 (Effective-margin calibration). *Under A3–A4 of Assumption 3.1, every component labeling $y \in [K]^{\mathcal{V}}$ satisfies*

$$\mathcal{J}(y) - \mathcal{J}(c) \geq \kappa_{\mathrm{sup}} M(y) - B_-(y) \geq \Delta_{\mathrm{id}} M(y). \tag{13}$$

Proof. For a class k, let $\mathcal{M}_k(y) = \{u \in \mathcal{V}_k : y_u \neq k\}$ and decompose the induced within-class graph on $\mathcal{M}_k(y)$ into connected sets D. Each within-class edge from D to $\mathcal{V}_k \setminus D$ is uncut by c and cut by y, contributing $\partial_k(D)$. On $u \in D$, the anchor increment is

$$\lambda\Big(\sum_{j \neq y_u} a_u^j - \sum_{j \neq k} a_u^j\Big) = \lambda(a_u^k - a_u^{y_u}).$$

Thus correct anchors contribute $+\lambda A_k^{\mathrm{corr}}(D)$, while all negative wrong-anchor contributions are contained in $-\lambda A_{\mathrm{wrong}}(y)$. Cross-semantic graph edges can reduce the cut objective only when a formerly cut edge becomes uncut; their total possible gain is $-\Lambda_{\mathrm{cross}}(y)$. Hence

$$\mathcal{J}(y) - \mathcal{J}(c) \geq \sum_k \sum_D [\partial_k(D) + \lambda A_k^{\mathrm{corr}}(D)] - B_-(y).$$

A3 lower-bounds every bracket by $\kappa_{\mathrm{sup}} m(D)$. The sets D partition the mismatched components, so their masses sum to $M(y)$. Finally apply Lemma A.2 and $\Delta_{\mathrm{id}} = \kappa_{\mathrm{sup}} - \rho_{\mathrm{leak}}$. □

A.4 Finite-sample semantic identification

We now pass from the population objective to its empirical counterpart. For a fixed component system, the excess empirical objective is supported only on components whose candidate labels differ from c. Its variance therefore scales with the mismatch mass rather than with a global constant. This localization yields the $1/m$ mismatch rate and the exact-recovery threshold used in Theorem 3.2. Assume first that the true component partition is fixed. Define the membership map $\omega(x) = u$ for $x \in C_u$ and $\omega(x) = \perp$ for $x \in C_0$. For $y \in [K]^{\mathcal{V}}$ write $\bar{y}(x) = y_{\omega(x)}$ on the resolved region. For $U_m = (X_1, \dots, X_m) \sim \mu^m$, let

$$\widehat{\mathcal{J}}_m^{\mathrm{cut}}(y) := \frac{1}{\binom{m}{2}} \sum_{i<\ell} \frac{1}{2} W_s(X_i, X_\ell) \mathbf{1}\{\omega(X_i), \omega(X_\ell) \in \mathcal{V},\ \bar{y}(X_i) \neq \bar{y}(X_\ell)\}, \tag{14}$$

$$\widehat{\mathcal{J}}_m^{\mathrm{anc}}(y) := \frac{\lambda}{m} \sum_{i=1}^{m} \mathbf{1}\{\omega(X_i) \in \mathcal{V}\} \sum_{j \neq \bar{y}(X_i)} \mathbf{1}\{X_i \in A_j^\tau(s)\}, \tag{15}$$

$$\widehat{\mathcal{J}}_m(y) := \widehat{\mathcal{J}}_m^{\mathrm{cut}}(y) + \widehat{\mathcal{J}}_m^{\mathrm{anc}}(y). \tag{16}$$

Then $\mathbb{E}\widehat{\mathcal{J}}_m(y) = \mathcal{J}(y)$. Let

$$\widehat{c}_m \in \arg\min_{y \in [K]^{\mathcal{V}}} \widehat{\mathcal{J}}_m(y). \tag{17}$$

For a fixed y, define $D(y) := \mathcal{J}(y) - \mathcal{J}(c)$ and $\widehat{D}_m(y) := \widehat{\mathcal{J}}_m(y) - \widehat{\mathcal{J}}_m(c)$.

Lemma A.5 (Localized concentration of the excess objective). *There is a universal $C > 0$ such that for every fixed y and $t > 0$,*

$$\Pr\left(|\widehat{D}_m(y) - D(y)| > C(1+\lambda)\left[\sqrt{\frac{M(y)t}{m}} + \frac{t}{m}\right]\right) \leq 2e^{-t}. \tag{18}$$

Proof. Let $R_y := \bigcup_{u:y_u \neq c(u)} C_u$, so $\mu(R_y) = M(y)$. For the anchor part define the one-sample excess increment

$$\xi_y(X) := \lambda \mathbf{1}\{\omega(X) \in \mathcal{V}\} \sum_{j=1}^{K} \big(\mathbf{1}\{j \neq \bar{y}(X)\} - \mathbf{1}\{j \neq \bar{c}(X)\}\big)\mathbf{1}\{X \in A_j^{\tau}(s)\}.$$

The anchor sets are disjoint, so $|\xi_y(X)| \leq \lambda$; if $X \notin R_y$ then $\xi_y(X) = 0$, hence $\mathbb{E}\xi_y(X)^2 \leq \lambda^2 M(y)$.

For the cut part define the symmetric order-two kernel

$$g_y(X, X') := \frac{1}{2} W_s(X, X')\big(\mathbf{1}\{\bar{y}(X) \neq \bar{y}(X')\} - \mathbf{1}\{\bar{c}(X) \neq \bar{c}(X')\}\big),$$

with points in C_0 contributing zero. Since $W_s \in [0, 1]$, $|g_y| \leq 1/2$. If neither endpoint lies in R_y, the two cut indicators coincide; therefore $\mathbb{E}g_y(X, X')^2 \leq \frac{1}{2} M(y)$.

Combine the two pieces into

$$H_y(X, X') := g_y(X, X') + \frac{1}{2}[\xi_y(X) + \xi_y(X')].$$

The empirical excess is the order-two U-statistic generated by H_y. Moreover, $H_y = 0$ unless at least one endpoint lies in R_y, $|H_y| \leq 1 + \lambda$, and

$$\mathrm{Var}(H_y(X, X')) \leq 2(1 + \lambda)^2 M(y).$$

Let $q = \lfloor m/2 \rfloor$. Hoeffding's permutation representation writes the U-statistic as the average, over random permutations, of averages of q independent pairs. Bernstein's moment-generating-function bound therefore applies to each independent-pair average, and Jensen's inequality transfers the same bound to the U-statistic. Its variance proxy is $O((1 + \lambda)^2 M(y)/q)$ and its scale is $O((1 + \lambda)/q)$. Chernoff's method for both tails, followed by $q \asymp m$, gives (18). □

Corollary A.6 (Uniform localized event)**.** *For any finite $\mathcal{C} \subseteq [K]^{\mathcal{V}}$, with probability at least $1 - \delta$, simultaneously for all $y \in \mathcal{C}$,*

$$|\widehat{D}_m(y) - D(y)| \leq C(1 + \lambda)\left[\sqrt{\frac{M(y)L_{\mathcal{C},\delta}}{m}} + \frac{L_{\mathcal{C},\delta}}{m}\right], \quad L_{\mathcal{C},\delta} := \log\frac{4|\mathcal{C}|}{\delta}. \tag{19}$$

Proof. Apply Lemma A.5 with $t = L_{\mathcal{C},\delta}$ and union bound over $\mathcal{C}$. □

Theorem A.7 (Fast identification rate)**.** *Assume $\Delta_{\mathrm{id}} > 0$ and let $\widehat{y}_{\mathcal{C},m} \in \arg\min_{y \in \mathcal{C}} \widehat{\mathcal{J}}_m(y)$ for a finite nonempty $\mathcal{C} \subseteq [K]^{\mathcal{V}}$. With probability at least $1 - \delta$, for every comparator $y^{\dagger} \in \mathcal{C}$,*

$$\begin{aligned} M(\widehat{y}_{\mathcal{C},m}) \leq \frac{2\Delta_{\dagger}}{\Delta_{\mathrm{id}}} + \frac{C(1 + \lambda)}{\Delta_{\mathrm{id}}}\sqrt{\frac{M(y^{\dagger})L_{\mathcal{C},\delta}}{m}} \\ + \frac{C'(1 + \lambda)^2 L_{\mathcal{C},\delta}}{\Delta_{\mathrm{id}}^2 m}, \end{aligned} \tag{20}$$

where $\Delta_{\dagger} = [\mathcal{J}(y^{\dagger}) - \mathcal{J}(c)]_+$ and C, C' are universal constants.

Proof. Work on the event in Corollary A.6. Empirical optimality gives $\widehat{D}_m(\widehat{y}_{\mathcal{C},m}) \leq \widehat{D}_m(y^{\dagger})$. Applying the uniform deviation bound on both sides and then Proposition A.4 yields, with $x = M(\widehat{y}_{\mathcal{C},m})$ and $A = C(1 + \lambda)\sqrt{L_{\mathcal{C},\delta}/m}$,

$$\Delta_{\mathrm{id}} x \leq \Delta_{\dagger} + A\sqrt{x} + A\sqrt{M(y^{\dagger})} + C(1 + \lambda)\frac{L_{\mathcal{C},\delta}}{m}.$$

Use Young's inequality $A\sqrt{x} \leq (\Delta_{\mathrm{id}}/2)x + A^2/(2\Delta_{\mathrm{id}})$, move the self term to the left, and absorb universal constants. □

Taking $\mathcal{C} = [K]^{\mathcal{V}}$ and $y^\dagger = c$ gives

$$M(\widehat{c}_m) \le E_{\mathrm{id}}(m,\delta) := \frac{C(1+\lambda)^2}{\Delta_{\mathrm{id}}^2 m}\big(V\log K + \log(4/\delta)\big). \tag{21}$$

If $m_{\min} := \min_u m_u$ and

$$m \ge \frac{2C(1+\lambda)^2}{\Delta_{\mathrm{id}}^2 m_{\min}}\big(V\log K + \log(4/\delta)\big), \tag{22}$$

then $M(\widehat{c}_m) < m_{\min}$; because every nonzero mismatch contains at least one component, this forces exact semantic identification $M(\widehat{c}_m) = 0$.

A.5 Projection onto the protocol-reachable class

The identification problem is formulated over the full labeling space $[K]^V$, whereas a fixed adaptation protocol Γ may realize only $\mathcal{Y}_\Gamma(C)$. We therefore project the identified labeling onto the protocol-admissible class and quantify the additional error due to empirical component masses and imperfect semantic identification. The identifier $\widehat{c}_m$ is computed over the full component-labeling space and need not be reachable by Γ. Define empirical component masses $\widehat{m}_u = m^{-1}\sum_i \mathbf{1}\{X_i \in C_u\}$ and, for an identified labeling $\widetilde{c}$, the empirical projection objective

$$\widehat{M}_{\widetilde{c}}(y) := \sum_{u\in\mathcal{V}} \widehat{m}_u \mathbf{1}\{y_u \ne \widetilde{c}_u\}, \qquad y \in \mathcal{Y}_\Gamma(C). \tag{23}$$

Let $\widehat{y}_m^{\mathrm{proj}} \in \arg\min_{y\in\mathcal{Y}_\Gamma(C)} \widehat{M}_{\widehat{c}_m}(y)$. Set

$$\Delta_m^{\mathrm{mass}}(\delta) := \sqrt{\frac{\log(2V/\delta)}{2m}}. \tag{24}$$

Lemma A.8 (Projection accuracy)**.** *With probability at least* $1-\delta$, $\max_u |\widehat{m}_u - m_u| \le \Delta_m^{\mathrm{mass}}(\delta)$, *and on this event for every* $\widetilde{c}$,

$$\sup_{y\in\mathcal{Y}_\Gamma(C)} |\widehat{M}_{\widetilde{c}}(y) - M(y)| \le V\Delta_m^{\mathrm{mass}}(\delta) + M(\widetilde{c}). \tag{25}$$

Proof. Hoeffding's inequality and a union bound give the mass event. For any reachable y,

$$\begin{aligned}|\widehat{M}_{\widetilde{c}}(y) - M(y)| \le \sum_u |\widehat{m}_u - m_u|\mathbf{1}\{y_u \ne \widetilde{c}_u\} \\ + \sum_u m_u|\mathbf{1}\{y_u \ne \widetilde{c}_u\} - \mathbf{1}\{y_u \ne c(u)\}|.\end{aligned}$$

The first term is at most $V\Delta_m^{\mathrm{mass}}$; the indicators in the second can differ only if $\widetilde{c}_u \ne c(u)$, giving $M(\widetilde{c})$. □

Theorem A.9 (Fixed-partition component-resolution PAC bound)**.** *Assume* $\Delta_{\mathrm{id}} > 0$. *With probability at least* $1-\delta$,

$$L_Q(h_{\widehat{y}_m^{\mathrm{proj}}}) \le L_{Q,C}^{\Gamma,\star} + \zeta + 2\alpha_{\mathrm{tot}} + 2V\Delta_m^{\mathrm{mass}}(\delta/2) + 2E_{\mathrm{id}}(m,\delta/2). \tag{26}$$

Proof. On the identification event, $M(\widehat{c}_m) \le E_{\mathrm{id}}(m,\delta/2)$. On the mass-concentration event, let $y^\circ \in \arg\min_{y\in\mathcal{Y}_\Gamma(C)} M(y)$, so $M(y^\circ) = \beta_\Gamma(C)$. Projection optimality and Lemma A.8 give

$$M(\widehat{y}_m^{\mathrm{proj}}) \le \beta_\Gamma(C) + 2V\Delta_m^{\mathrm{mass}}(\delta/2) + 2M(\widehat{c}_m).$$

Apply Lemma A.1 and $\beta_\Gamma(C) \le L_{Q,C}^{\Gamma,\star} + \alpha_{\mathrm{tot}}$, then union bound the two events. □

A.6 Estimated components and the complete PAC bound

The previous results condition on the population component system C. Under A2, the learner instead recovers an estimated system $\widehat{C}$ from an independent unlabeled split. Lemma A.11 controls the perturbation of component masses, anchor masses, graph weights, and the effective identification margin under the matched symmetric-difference error δ_C. protocol stability then transfers protocol-admissible labelings from C to $\widehat{C}$.

Definition A.10 (Pointwise consistent component recovery)**.** A single label-free estimator $\mathcal{A}_0(s, D_1)$ is pointwise consistent if, for every fixed admissible pair (Q, s) and every $\delta \in (0, 1)$, there exist a finite burn-in $m_{\mathrm{part}}(Q, s, \delta)$ and a deterministic error bound $\delta_C(m_1, \delta; Q, s) \to 0$ as $m_1 \to \infty$ such that, for every $m_1 \geq m_{\mathrm{part}}(Q, s, \delta)$, with probability at least $1 - \delta$, there is a bijection $\pi : \mathcal{V} \to \widehat{\mathcal{V}}$ satisfying

$$\max_{u \in \mathcal{V}} \mu(C_u \triangle \widehat{C}_{\pi(u)}) \leq \delta_C(m_1, \delta; Q, s). \tag{27}$$

The estimator does not depend on the unknown Q; only its burn-in and error bound may depend on Q.

Throughout the remaining proofs, (Q, s) is fixed and we abbreviate $\delta_C(m_1, \delta; Q, s)$ as $\delta_C(m_1, \delta)$. Theorem 3.2 uses only this pointwise consistency interface and does not require a recovery rate shared by all admissible target distributions.

We next quantify the stability of the structural quantities under component-recovery error. Suppose $\delta_C := \max_u \mu(C_u \triangle \widehat{C}_{\pi(u)}) \leq m_{\min}/2$. Equip the estimated components with labels $\widehat{c} = c \circ \pi^{-1}$ and compute the population masses, anchors, and weights on the estimated cells.

Lemma A.11 (Structural transfer under component perturbation)**.** *Let*

$$C_{\mathrm{str}} := \frac{4(2V + \lambda + \kappa_{\mathrm{sup}})}{m_{\min}}. \tag{28}$$

Then the estimated semantic system satisfies

$$\widehat{m}^{\mathrm{pop}}_{\min} \geq m_{\min}/2, \tag{29}$$

$$\widehat{\zeta}^{\mathrm{pop}} \leq \zeta + V\delta_C, \qquad \widehat{\alpha}^{\mathrm{pop}}_{\mathrm{tot}} \leq \alpha_{\mathrm{tot}} + V\delta_C, \tag{30}$$

$$\widehat{\kappa}^{\mathrm{pop}}_{\mathrm{sup}} \geq \kappa_{\mathrm{sup}} - \tfrac{1}{2}C_{\mathrm{str}}\delta_C, \qquad \widehat{\rho}^{\mathrm{pop}}_{\mathrm{leak}} \leq \rho_{\mathrm{leak}} + \tfrac{1}{2}C_{\mathrm{str}}\delta_C, \tag{31}$$

$$\widehat{\Delta}^{\mathrm{pop}}_{\mathrm{id}} \geq \Delta_{\mathrm{id}} - C_{\mathrm{str}}\delta_C. \tag{32}$$

Moreover, for every component labeling y transported by π, its mismatch mass changes by at most $V\delta_C$. Under protocol stability, the image of every true reachable labeling is reachable on the estimated components, so $\widehat{\beta}^{\mathrm{pop}}_{\Gamma} \leq \beta_{\Gamma}(C) + V\delta_C$.

Proof. Mass and anchor terms change by at most δ_C per matched cell. Since $W_s \in [0, 1]$, for matched cells C_u, C_v and $\widehat{C}_{\pi(u)}, \widehat{C}_{\pi(v)}$,

$$|w_{uv} - \widehat{w}_{\pi(u)\pi(v)}| \leq \mu(C_u \triangle \widehat{C}_{\pi(u)}) + \mu(C_v \triangle \widehat{C}_{\pi(v)}) \leq 2\delta_C.$$

Hence the support numerator of a set S changes by at most $|S|[2V + \lambda]\delta_C$, while its mass changes by at most $|S|\delta_C$. By Lemma A.3, it is enough to control all nonempty subsets. Because every estimated cell has mass at least $m_{\min}/2$, division yields the support loss in (31) with the displayed conservative constant. The leakage bound is the analogous one-cell calculation, again using the $m_{\min}/2$ denominator. Subtracting the two proves (32). Mismatch changes only through cell-mass perturbations, giving $V\delta_C$; evaluating the estimated projection error at the transported true minimizer gives the final claim. □

Theorem A.12 (Recover–identify–project bound)**.** *Split the sample into independent D_1, D_2 of sizes m_1, m_2. On D_1, recover $\widehat{C} = \mathcal{A}_0(s, D_1)$. On D_2, identify over the full estimated component-labeling space and project to $\mathcal{Y}_{\Gamma}(\widehat{C})$ as in the main text. Suppose*

$$\delta_C(m_1, \delta/3) \leq \min\left\{\frac{m_{\min}}{2}, \frac{\Delta_{\mathrm{id}}}{2C_{\mathrm{str}}}\right\}. \tag{33}$$

Then with probability at least $1-\delta$*,*

$$L_Q(\widehat{h}) \le L_{Q,C}^{\Gamma,\star} + \zeta + 2\alpha_{\mathrm{tot}} + 2V\Delta_{m_2}^{\mathrm{mass}}(\delta/3) + \frac{C(1+\lambda)^2[V\log K + \log(12/\delta)]}{(\Delta_{\mathrm{id}} - C_{\mathrm{str}}\delta_C)^2 m_2} + 3V\delta_C, \tag{34}$$

where $\delta_C = \delta_C(m_1, \delta/3)$.

Proof. The recovery event on D_1 holds with probability at least $1-\delta/3$. Condition on it. The estimated partition is now fixed and D_2 is still i.i.d. from μ. Lemma A.11 yields an estimated component system with effective margin at least $\Delta_{\mathrm{id}} - C_{\mathrm{str}}\delta_C \ge \Delta_{\mathrm{id}}/2$. Apply Theorem A.9 to this estimated system using the two remaining failure budgets $\delta/3$. Finally use (30) and the transported reachable-class bound to replace the estimated structural resolution terms by their true counterparts, incurring at most $3V\delta_C$. A union bound completes the proof. □

Taking $m_1 = \lfloor m/2 \rfloor$ and $m_2 = m - m_1$ in Theorem A.12 gives the big-O statement in Theorem 3.2.

Implication for pointwise learnability. Fix an admissible pair (Q, s) and $\varepsilon, \delta \in (0,1)$. For this fixed problem, $V < \infty$, $m_{\min} > 0$, $\Delta_{\mathrm{id}} > 0$, and $C_{\mathrm{str}} < \infty$ are fixed constants. Pointwise component consistency implies that $\delta_C(m_1, \delta/3) \to 0$, so (33) holds for all sufficiently large m. The effective margin is then at least $\Delta_{\mathrm{id}}/2$, and every remainder term in (34) tends to zero. Hence there exists a finite $m_0(Q, s, \varepsilon, \delta)$ such that (2) holds for every $m \ge m_0(Q, s, \varepsilon, \delta)$. The learner uses the same fixed sample split and the same recover–identify–project procedure for every admissible Q. Only the sample threshold depends on the realized target distribution.

A.7 Necessity of the structural ingredients

Theorem 3.2 is a sufficiency result. The constructions below show that each of A1–A4 and protocol stability excludes a distinct obstruction. They are leave-one counterexamples: after removing one structural role, while allowing no alternative assumption to replace it, the enlarged model class contains observationally indistinguishable target worlds or a non-vanishing protocol-dependent error floor. These results are not intended as an if-and-only-if characterization of TTA-PAC learnability.

	Role in Theorem 3.2	Missing structure	Consequence
A1	Controls semantic resolution	Labels may vary arbitrarily inside a resolved component	Unavoidable risk on an observationally hidden subregion
A2	Marks which component scale is semantic	Two worlds can share every unlabeled/source observable but use different semantic partitions	The semantic decomposition itself is not identifiable
A3	Supplies source-oriented support	A positive-mass semantic block can have neither boundary support nor a correct anchor	The whole block can be label-flipped without an observable cue
A4	Requires support to dominate leakage	Correct support can be matched or exceeded by wrong-oriented evidence	Opposite semantic labels remain observationally ambiguous
protocol stability	Connects semantics to the deployed protocol	The statistically correct component labels may be unreachable	A protocol-dependent risk floor remains even after perfect identification

The form of the obstruction differs across the five assumptions. A2 and A4 lead directly to indistinguishable target worlds requiring incompatible semantic decisions. A1 controls the resolution error that remains invisible within a component; A3 rules out a semantic block with no source-oriented support; and protocol stability rules out statistically identifiable but protocol-unreachable decisions. The perturbation-stability clause associated with protocol stability is used only to transfer the fixed-partition guarantee to $\widehat{C}$ and is not claimed to be information-theoretically necessary.

Proposition A.13 (A1: no guarantee below an uncontrolled component resolution)**.** *Suppose the admissible family places no uniform control on label variation inside a resolved component. Then for every* $\beta > 0$ *allowed as an uncontrolled within-component region, there are two target worlds with*

the same unlabeled marginal and the same source-induced observables whose labels are opposite on a measurable set R of mass β. Consequently every unlabeled learner has target risk at least $\beta/2$ in one of the two worlds. Thus a uniform low-risk statement requires a component-resolution condition such as near-purity (or another assumption that controls the same unresolved label variation).

Proof. Fix a component C_u containing a measurable set R of mass $\mu(R) = \beta$. Keep μ, the source model, confidence scores, representation, affinity, anchors, and protocol fixed. Construct Q and Q' that agree outside R and assign opposite deterministic labels on R. For every fixed classifier h, the two losses restricted to R sum to β; hence $L_Q(h) + L_{Q'}(h) \geq \beta$. Since an unlabeled learner has the same output distribution in the two worlds, one world has risk at least $\beta/2$ with probability at least $1/2$. Near-purity is precisely what converts this otherwise uncontrolled resolution error into the explicit α_{tot} term appearing in the theorem. □

Failure without A2. The following construction keeps the components pure and preserves positive source support and a strictly positive margin $\Delta_{\mathrm{id}} = 1/4$ in both worlds. The worlds differ only in which semantic partition is valid, while all quantities available to an unlabeled learner have the same distribution. Thus existence of a favorable partition is insufficient unless its scale is recoverable from observable information.

Proposition A.14 (Non-identifiability of an unmarked semantic scale)**.** *Near-purity, positive support, and positive support–leakage margin are not sufficient if the semantic partition is only asserted to exist. There are two observationally identical target worlds with the same μ, source model, confidences, representation, affinity, and anchors, each admitting a pure component system with $\Delta_{\mathrm{id}} = 1/4$ and zero component-resolution oracle risk, but with different semantic component systems. For every unlabeled learner and every m, one of the two worlds has failure probability at least $1/2$ at target risk $1/4$.*

Proof. Let $K = 2$, set $\lambda = 1$, and let $\mathcal{X} = A_1 \sqcup A_2$ have masses $1/2, 1/2$. Take source affinity 0.9 within each A_i and 0.5 across them. Put a class-1 anchor on all of A_1 and a class-2 anchor on a subset of A_2 of mass $1/4$. Let the reachable class contain the constant class-1 classifier and the classifier that labels A_1 as 1 and A_2 as 2.

In world Q, the semantic components are A_1, A_2 with deterministic labels 1 and 2. The singleton support ratios are at least $1/2$ and the cross-edge leakage rate is $1/4$, hence $\Delta_{\mathrm{id}} = 1/4$. In world Q', the single semantic component is $A_1 \cup A_2$ with deterministic label 1; its correct anchor ratio is again $1/2$ and the class-2 anchor now contributes leakage $1/4$, again giving $\Delta_{\mathrm{id}} = 1/4$. The component-resolution oracle risk is zero in both worlds. Yet all unlabeled observations have the same law. For every fixed output h, $L_Q(h) + L_{Q'}(h) \geq \mu(A_2) = 1/2$ because the two worlds disagree on all of A_2. The two failure events at risk $1/4$ therefore cover the common output space; at least one has probability at least $1/2$. □

Proposition A.15 (A3: an unsupported semantic block can be flipped)**.** *Suppose a fixed semantic component system contains a nonempty connected block $S \subseteq \mathcal{V}_k$ of mass $m(S) = \beta > 0$ with $\partial_k(S) = 0$ and $A_k^{\mathrm{corr}}(S) = 0$. If the admissible family contains the corresponding label-flipped world in which the semantic label of all components in S is changed from k to some $\ell \neq k$ while all unlabeled source-induced objects are unchanged, then every unlabeled learner has risk at least $\beta/2 - \alpha_{\mathrm{tot}}$ in one of the two worlds; for pure S, the lower bound is $\beta/2$.*

Proof. The block has neither a same-class boundary signal nor correctly oriented anchor mass, so the source-induced observations used by the learner can be kept fixed while its semantic label is flipped. Let a common output incur risk r on S in the first world. In the flipped world its risk on S is at least $\beta - r$, up to the total minority-label mass allowed on that block. Hence the larger of the two risks is at least $\max\{r, \beta - r\} - \alpha_{\mathrm{tot}} \geq \beta/2 - \alpha_{\mathrm{tot}}$. □

Failure without A4. A3 guarantees the presence of source-oriented support but does not control competing evidence. The construction below preserves positive support while introducing symmetric wrong anchors with $\rho_{\mathrm{leak}} \geq \kappa_{\mathrm{sup}}$, so that $\Delta_{\mathrm{id}} \leq 0$. The resulting target worlds are observationally identical but have opposite optimal component labels.

Theorem A.16 (Ambiguity without a positive support–leakage margin)**.** *There is a one-component binary family satisfying positive source support but with $\Delta_{\mathrm{id}} \le 0$ for which no unlabeled learner achieves additive component-resolution PAC learning.*

Proof. Take one component of mass m_u, no graph edges, and symmetric source anchor masses $a_u^1 = a_u^2 = a$ with $\lambda a \ge \kappa_{\mathrm{sup}} m_u$. Construct two worlds with the same marginal and source object but conditional class-1 probabilities $1/2+\gamma$ and $1/2-\gamma$. The optimal component labels are opposite and choosing the wrong label costs $2\gamma m_u$. Unlabeled observations are identical, so a common decision fails in one world with probability at least $1/2$. Here the wrong-anchor leakage rate is $\rho_{\mathrm{leak}} = \lambda a / m_u \ge \kappa_{\mathrm{sup}}$, so $\Delta_{\mathrm{id}} \le 0$. □

Thus source-oriented support alone does not identify the semantic label when wrong-oriented evidence can match or exceed it.

Proposition A.17 (Protocol stability: protocol incompatibility creates irreducible semantic error)**.** *Define the semantic mismatch of the best protocol-reachable component labeling by*

$$\beta_\Gamma(C) := \inf_{y \in \mathcal{Y}_\Gamma(C)} M(y).$$

If $\beta_\Gamma(C) > 0$, then every protocol-reachable component-constant output satisfies

$$L_Q(h_y) \ge \beta_\Gamma(C) - \alpha_{\mathrm{tot}}.$$

Thus recovering the semantic labels is not enough to obtain a low-risk TTA predictor unless the protocol can realize (or faithfully project to) those semantic decisions.

Proof. Lemma A.1 gives $L_Q(h_y) \ge M(y) - \alpha_{\mathrm{tot}}$ for every reachable labeling y. Minimize over $\mathcal{Y}_\Gamma(C)$. □

Remark A.18 (Scope of the necessity statements). A2 and A4 exclude observationally indistinguishable target worlds requiring incompatible semantic decisions. These obstructions also apply to pointwise learnability: if a single learner satisfied the guarantee on both members of a fixed indistinguishable pair, then any sample size exceeding both distribution-dependent thresholds would have to satisfy both guarantees, contradicting the two-world lower bound.

A1 controls the unresolved label variation accounted for by the residual error, A3 excludes unsupported semantic blocks, and protocol compatibility controls the attainable semantic error. The perturbation-stability condition is used to transfer the fixed-partition guarantee to estimated components; it is not claimed to be information-theoretically necessary. These results establish distinct structural obstructions rather than an if-and-only-if characterization of pointwise component-resolution learnability.

Necessity of the component-resolution comparator. The structural assumptions identify the dominant label of each resolved component. They do not reveal arbitrary label variation inside a component. Accordingly, an unrestricted reachable oracle can be strictly stronger than any comparator that is identifiable from unlabeled data. The following two-world construction formalizes this gap.

Theorem A.19 (Unrestricted reachable oracle cannot be learned uniformly)**.** *Fix any finite $\beta \ge 1$. Suppose an admissible component C_u has minority-label budget at least $\alpha > 0$, and the reachable class can either label all of C_u by its majority class or separately flip a measurable subregion $R \subset C_u$ of mass α. Then no unlabeled learner can uniformly guarantee $L_Q(\widehat{h}) \le \beta \inf_{h \in \mathcal{H}_s^\Gamma} L_Q(h) + \varepsilon$ for $\varepsilon < \alpha/2$ and $\delta < 1/2$.*

Proof. Construct two worlds with identical unlabeled marginal and source objects. In the first, R has the minority label and the rest of the component has the majority label; in the second, the entire component has the majority label. The two reachable hypotheses described in the theorem have zero risk in the two respective worlds, so the unrestricted oracle risk is zero in both. For any common output h, the two true labelings disagree on R, hence $L_Q(h) + L_{Q'}(h) \ge \alpha$. Because the learner's output distribution is identical in the two worlds, one world has risk at least $\alpha/2$ with probability at least $1/2$. Multiplying a zero oracle risk by finite β does not change the contradiction. □

Therefore the component-resolution reachable oracle is the appropriate comparator for a uniform unlabeled guarantee: when the unrestricted oracle has zero risk in both worlds, no finite multiplicative relaxation removes the ambiguity below the component resolution.

A.8 Lower bound for exact semantic recovery

We next show that the leading dependence in the exact-recovery threshold is not an artifact of the proof technique. A binary Assouad construction gives a lower bound of order $\Delta_{\mathrm{id}}^{-2} m_{\min}^{-1}$, up to logarithmic and labeling-complexity factors. Consider binary labels and components with no graph edges. For a sign vector $\sigma \in \{\pm 1\}^V$, let $c_\sigma(u)$ denote the corresponding semantic label of component u, and define $M_\sigma(y) := \sum_u m_u \mathbf{1}\{y_u \neq c_\sigma(u)\}$. Inside component u, let two observable source-anchor regions have total mass fraction a but an unknown bias sign $\sigma_u \in \{\pm 1\}$:

$$\mu_\sigma(A_u^{\sigma_u}) = m_u \frac{a(1+\gamma)}{2}, \qquad \mu_\sigma(A_u^{-\sigma_u}) = m_u \frac{a(1-\gamma)}{2}. \tag{35}$$

The semantic component label is the sign σ_u. Then $\kappa_{\mathrm{sup}} = \lambda a(1+\gamma)/2$, $\rho_{\mathrm{leak}} = \lambda a(1-\gamma)/2$, and $\Delta_{\mathrm{id}} = \lambda a \gamma$.

Theorem A.20 (Assouad lower bound for exact component recovery)**.** *For every learner $\widehat{y}(U_m)$ in the family (35),*

$$\sup_\sigma \mathbb{E}_\sigma M_\sigma(\widehat{y}) \geq \frac{1}{2} \sum_u m_u \left(1 - \sqrt{2 m m_u a \gamma^2}\right)_+ . \tag{36}$$

Consequently, for $a, \lambda = \Theta(1)$ and any fixed $\delta \in (0, 1/2)$, exact recovery of all component labels with probability at least $1 - \delta$ throughout this hard family requires, in the worst case,

$$m = \Omega\left(\frac{1}{\Delta_{\mathrm{id}}^2 m_{\min}}\right). \tag{37}$$

Proof. If two sign vectors differ only at coordinate u, only the two anchor probabilities in C_u change. For $\gamma \leq 1/2$,

$$\mathrm{KL}(\mu_\sigma \| \mu_{\sigma^{(u)}}) \leq 4 m_u a \gamma^2,$$

so the m-sample KL is at most $4 m m_u a \gamma^2$. Pinsker therefore gives adjacent total variation at most $\sqrt{2 m m_u a \gamma^2}$. Apply Assouad's lemma to the hypercube with weighted Hamming loss $\sum_u m_u \mathbf{1}\{\widehat{\sigma}_u \neq \sigma_u\}$ to obtain (36). To obtain the exact-recovery lower bound, consider two adjacent worlds differing only on a component of mass $m_{\min}$. Any exact-recovery rule must distinguish these two worlds. The binary testing bound therefore gives a worst-case failure probability of at least $\frac{1}{2}(1 - \sqrt{2 m m_{\min} a \gamma^2})_+$. For this probability to be at most a fixed $\delta \in (0, 1/2)$, it is necessary that $m \geq (1-2\delta)^2 / (2 m_{\min} a \gamma^2)$. Since $\Delta_{\mathrm{id}} = \lambda a \gamma$ and $a, \lambda = \Theta(1)$, this yields (37). □

Relation to pointwise learnability. This is a worst-case finite-sample lower bound over the hard family, not an impossibility result for a fixed positive-margin problem. For each fixed admissible Q, the identification margin and component masses are positive constants, and the upper bound permits recovery with sufficiently many samples. The lower bound explains why the required sample size can grow without bound as the identification margin decreases, or, for exact recovery, as the smallest component mass decreases. This is compatible with the distribution-dependent threshold in (2).

A.9 Concrete sufficient certificates for the structural conditions

The PAC theorem requires only the abstract conditions A2–A4. For completeness, we give two sufficient constructions. The first makes the semantic scale observable from unlabeled source-induced affinities and hence provides one realization of A2. The second gives a spectral lower bound on κ_{sup} and therefore a polynomial-time sufficient certificate for A3–A4. Neither construction is needed for the preceding PAC proof.

A.9.1 AN OBSERVABLE MERGE-GAP CONDITION FOR A2

Let $A_1, \dots, A_{N_\star}$ be measurable atoms of positive mass whose unions form the semantic components. For $r \neq r'$, let $\bar{s}_{rr'} = \operatorname{ess\,sup}_{A_r \times A_{r'}} W_s$ and let $\mathcal{T}_\star$ be a maximum spanning tree on the atom graph with weights $\bar{s}_{rr'}$. Let $H_\star$ be its edge-weight set together with sentinels $\{0, 1\}$.

Condition A.21 (Observable dominant merge gap)**.** Assume $\zeta = 0$ and there is a distinguished maximal gap $G_\star = (a_\star, b_\star)$ of $H_\star$, width $g_\star = b_\star - a_\star > 0$, and a radius $\eta_\star \leq g_\star/16$ such that:

(i) cutting $\mathcal{T}_\star$ at level $b_\star$ gives exactly the semantic atom groups, and every cross-semantic atom pair has essential supremum at most $a_\star$;

(ii) within each atom, $\operatorname{ess\,inf}_{A_r \times A_r} W_s \geq b_\star - \eta_\star$;

(iii) each population tree edge of weight h_e has positive pair mass q_e at similarity at least $h_e - \eta_\star$, and $q_\star := \min_e q_e > 0$;

(iv) every other maximal gap of $H_\star$ has width at most $g_\star/2$.

Let $p_{\min} := \min_r \mu(A_r) > 0$.

Given D_1, compute a maximum spanning tree of the complete sample affinity graph, choose its widest merge-height gap, cut at the midpoint, and extend the sample clusters to $\mathcal{X}$ by maximum similarity to a sampled cluster.

Proposition A.22 (Exact recovery from the dominant merge gap)**.** *Under Condition A.21, if*

$$m_1 \geq \max\left\{ \frac{1}{p_{\min}} \log \frac{2N_\star}{\delta}, \quad 2\frac{1}{q_\star} \log \frac{2\max\{N_\star - 1, 1\}}{\delta} + 3 \right\}, \tag{38}$$

then with probability at least $1 - \delta$ the estimator recovers all semantic components up to μ-null sets. In particular it satisfies A2 with $\delta_C = 0$ beyond the burn-in.

Proof. With probability at least $1 - \delta/2$, every atom is represented in the sample by the coupon-collector union bound. Fix a population tree edge e. Among $\lfloor m_1/2 \rfloor$ predetermined disjoint sample pairs, the chance that none lands in the near-maximal set defining q_e is at most $\exp[-q_e \lfloor m_1/2 \rfloor]$. The second term in (38) and a union bound over tree edges therefore ensure, with probability at least $1 - \delta/2$, that every critical merge edge is witnessed within $\eta_\star$ of its population height.

Work on the intersection. A maximum-spanning-tree bottleneck argument shows that every population merge height h_e has a sample merge height in $[h_e - \eta_\star, h_e]$. For every threshold $t \in (a_\star, b_\star - \eta_\star)$, atom tightness and the witnessed critical edges connect all sampled atoms within a semantic component, while the cross-semantic ceiling forbids any merge across semantic components. Thus the sample partition is constant and semantically correct throughout this whole interval. Its empirical persistence gap has width at least $g_\star - \eta_\star$. Any competing empirical gap can have width at most $g_\star/2 + 2\eta_\star$ by the one-sided localization of merge heights and the population dominance condition. Since $\eta_\star \leq g_\star/16$, $g_\star - \eta_\star > g_\star/2 + 2\eta_\star$, so the semantic gap is the unique widest empirical gap and is selected by the estimator.

Finally, every x in atom A_r has similarity at least $b_\star - \eta_\star$ to a sampled point from the same atom and at most $a_\star$ to every sampled point in another semantic component. The extension rule therefore assigns x to the correct component, except on null sets. □

A.9.2 A SPECTRAL SUFFICIENT CONDITION FOR A3–A4

For a class k, let $W_k = (w_{uv})_{u,v \in \mathcal{V}_k}$, L_k its graph Laplacian, $P_k = \lambda \operatorname{diag}(a_u^k)$, and $M_k = \operatorname{diag}(m_u)$. Define

$$\kappa_{\text{spec}} := \min_k \lambda_{\min}\big(M_k^{-1/2}(L_k + P_k)M_k^{-1/2}\big). \tag{39}$$

Proposition A.23 (Spectral lower certificate)**.** *For every class k and every nonempty $S \subseteq \mathcal{V}_k$,*

$$\partial_k(S) + \lambda A_k^{\text{corr}}(S) \geq \kappa_{\text{spec}} m(S). \tag{40}$$

Hence $\kappa_{\text{spec}} > \rho_{\text{leak}}$ is a polynomial-time sufficient certificate for a positive effective margin.

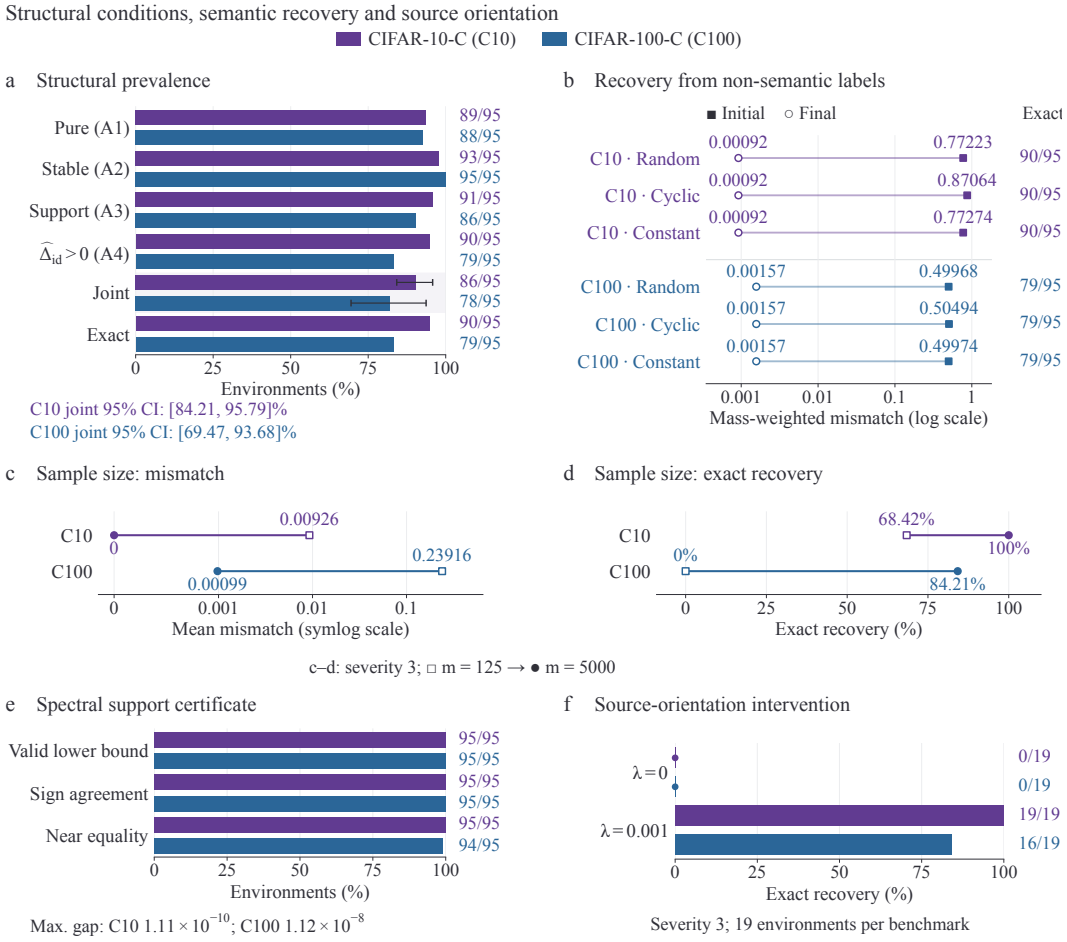

Figure 4: Structural conditions, semantic recovery, and source orientation. (a) Prevalence of A1–A4, their joint satisfaction, and exact recovery over 95 corruption–severity environments per benchmark. (b) Initial and final component-mass-weighted mismatch from random, cyclic, and constant initializations, with exact recovery counts at right. (c,d) Objective-size endpoints at severity 3. (e) Validity and sign agreement of the spectral certificate. (f) Exact recovery under the source-orientation intervention over 19 severity-3 environments. C10 and C100 denote CIFAR-10-C and CIFAR-100-C.

Proof. Let χ_S be the indicator of S. Then $\chi_S^\top L_k \chi_S = \partial_k(S)$, $\chi_S^\top P_k \chi_S = \lambda A_k^{\mathrm{corr}}(S)$, and $\chi_S^\top M_k \chi_S = m(S)$. With $z = M_k^{1/2}\chi_S$, the ratio is the Rayleigh quotient of $M_k^{-1/2}(L_k+P_k)M_k^{-1/2}$ and is therefore at least its smallest eigenvalue. Minimize over k. □

B Additional Experiments for PAC Learnability

This section gives the experimental details for Section 3.3. We evaluate A1–A4 on CIFAR-10-C/CIFAR-100-C (Hendrycks & Dietterich, 2019), test the finite-sample and source-orientation predictions, and study a separate label-free analogue of the effective identification margin. In the structural evaluation, target labels are revealed only after the component system, affinities, and anchors have been fixed.

B.1 Empirical Validation of Recoverable Semantic Structure

We empirically examine whether the structural conditions underlying Assumption 3.1 arise on natural distribution shifts using CIFAR-10-C and CIFAR-100-C (Hendrycks & Dietterich, 2019). The evaluation covers all 19 corruption types and five severity levels, giving 95 target environments per

benchmark. Each environment contains $10{,}000$ target examples. Both audits start from the same pretrained ResNet-50 backbone, but they deliberately instantiate the component-recovery interface differently. CIFAR-10-C uses the frozen source representation directly; CIFAR-100-C first performs an unlabeled AdaBN recalibration and then uses the resulting probabilities and features for label-free component estimation. For any fixed configuration, no target labels or gradient-based parameter updates are used to construct either component system. Within each run, target labels are withheld during component construction, graph construction, anchor assignment, and semantic identification, and are revealed only afterwards for evaluating the structural conditions and recovery accuracy.

The purpose of using two distinct component estimators is to test whether the same recoverable-semantic-structure conditions can be realized under different label-free resolutions of the target domain, rather than to tie the theory to one particular clustering algorithm.

Construction of the empirical component system. For each target environment, the $10{,}000$ examples are randomly divided into two equal subsets, D_1 and D_2. The role of this split is to separate *component discovery* from *semantic identification*: D_1 is used to fit a label-free component estimator and all quantities required to apply it to unseen target samples, after which the estimator is frozen and transferred to the independent D_2.

CIFAR-10-C. We construct components from the frozen source-model representation and predictions using prediction-constrained mutual-kNN. Neighborhood connections are permitted only between samples receiving the same source top-1 prediction, while source-feature geometry determines the finer connected components. Samples that cannot be assigned to a supported component are placed in the unresolved region C_0.

CIFAR-100-C. We intentionally use a different label-free construction. Batch-normalization statistics are first recalibrated from unlabeled target inputs (AdaBN), and the resulting target probabilities and features are used to fit a label-free clustering estimator on D_1. The fitted estimator is then frozen and applied to D_2. In particular, the CIFAR-100-C components are *not* generated by the prediction-constrained mutual-kNN procedure used on CIFAR-10-C. This second construction provides an independent test of the component-recovery interface under a substantially different unlabeled partitioning mechanism.

Once the dataset-specific estimator has been fitted, its D_2 assignments define the empirical component masses and the graph used for semantic identification. Pairwise affinities are approximated with 128 Gaussian random Fourier features; the bandwidth multiplier is 0.1 on CIFAR-10-C and 0.05 on CIFAR-100-C. Source anchors use the common confidence threshold $\tau = 0.7$ and graph–anchor coefficient $\lambda = 0.05$. The mean anchor coverages are 63.22% on CIFAR-10-C and 70.74% on CIFAR-100-C. The reported dataset-level configurations were chosen during exploratory development using post-hoc structural audit outcomes. For the main audit, each selected configuration is then applied unchanged across the corresponding 95 environments. Within each run, target labels are revealed only after the learner-visible component system, graph, and anchors have been frozen. Thus, label use occurs in the outer research-stage configuration selection and the post-hoc audit, not in the learner's execution.

Empirical realization of A1–A4. We next associate the first four conditions of Assumption 3.1 with observable properties of the constructed component system. For this audit only, target labels are revealed after the partition, graph, and anchors have been fixed.

A1 requires the resolved components to be nearly class-pure. Let $c(u)$ be the majority target class in component C_u and let α_u denote its minority-label fraction. Consistent with the notation of Appendix A, we use the total minority-label mass $\alpha_{\mathrm{tot}} = \sum_u m_u \alpha_u$ and regard A1 as satisfied when $\alpha_{\mathrm{tot}} \leq 0.25$. Thus the "Pure" criterion measures whether replacing each component by a single semantic label incurs only a small resolution error.

A2 concerns a different property: the component system must be recoverable from unlabeled target observations, rather than depending sensitively on the particular examples used to fit it. We therefore repeat the complete component-recovery procedure five times using five independent random D_1/D_2 half-splits of the same target environment. Each run fits a new component estimator from its own D_1, without using target labels. To make the resulting partitions directly comparable, all five fitted estimators are then applied to the same full set of $10{,}000$ target examples. This produces five

independently estimated partitions of an identical reference set. We call the component structure "Stable" when their pairwise normalized variation of information is at most 0.25 and the recovered component counts remain consistent across the five fits. This is an empirical counterpart of the consistency requirement in A2: it measures the stability of the unlabeled partition itself, not semantic agreement with the target labels.

A3 and A4 concern semantic identification once the component system has been fixed. Using the post-hoc component labels $c(u)$ only for evaluation, we compute the empirical source-oriented support constant $\widehat{\kappa}_{\mathrm{sup}}$ defined by the graph and anchors. The "Support" criterion corresponds to A3 and requires positive source-oriented support for every semantic block. Importantly, A3 is not audited by enumerating a sampled collection of connected sets. For each semantic class, the minimum support-to-mass ratio over all nonempty component subsets is computed exactly by a root-pinned fractional minimization: Dinkelbach iterations reduce each ratio problem to an s–t min-cut. By Lemma A.3, the minimum over all nonempty subsets equals the minimum over connected subsets required by A3. Thus the reported "Support" indicator audits the full setwise condition rather than a sampled proxy. We then compute the worst cross-semantic leakage $\widehat{\rho}_{\mathrm{leak}}$ and the empirical identification margin

$$\widehat{\Delta}_{\mathrm{id}} := \widehat{\kappa}_{\mathrm{sup}} - \widehat{\rho}_{\mathrm{leak}}.$$

A4 is satisfied when $\widehat{\Delta}_{\mathrm{id}} > 0$; numerically, values larger than 10^{-12} are treated as positive. "Joint" denotes simultaneous satisfaction of A1–A4. Protocol stability is deliberately excluded from this incidence analysis because it concerns whether an already identified component labeling can be realized by the protocol-reachable class $\mathcal{Y}_\Gamma(C)$, rather than whether the recoverable semantic structure exists in the target distribution.

The four conditions hold jointly in $86/95$ environments on CIFAR-10-C and $78/95$ environments on CIFAR-100-C. More importantly for the identification argument, the sign of $\widehat{\Delta}_{\mathrm{id}}$ agrees with exact semantic recovery in all 190 environments: the 90 CIFAR-10-C environments and 79 CIFAR-100-C environments with positive empirical identification margin are exactly those in which the semantic component labeling is recovered.

Recovery is not inherited from the initialization. To verify that the preceding result reflects the graph–anchor objective rather than a favorable starting labeling, we minimize the same empirical objective from several deliberately different initializations. The *source-majority* initialization assigns each component the majority top-1 prediction of the frozen source model. The *random* initialization assigns component labels uniformly at random. The *cyclic* initialization deliberately moves every source-majority label to the next class, $y_u^{(0)} = (y_u^{\mathrm{src}} + 1) \bmod K$, while the *constant* initialization assigns every component to class 0. None of these initializations uses target labels.

We optimize the multiclass Potts graph–anchor objective by alpha-expansion. For a class α, one expansion step solves a graph-cut subproblem that jointly determines which components should retain their current label and which should switch to α; expansion moves are repeated over the classes until no further decrease of the objective is obtained. Exact recovery means that the final label of every resolved component equals its post-hoc majority target class.

Despite starting from substantially different and intentionally misaligned labelings, the random, cyclic, and constant initializations converge to the same final mismatch and the same exact-recovery rate on both benchmarks. Together, the $169/169$ recovery in the certified regime, the $21/21$ failure outside it, and the agreement across deliberately misaligned initializations show that the observed threshold is not inherited from a favorable initialization. The positive-margin side is the expected optimization consequence of the calibration inequality; the non-positive side provides the complementary empirical separation.

Decomposing source improvement. We isolate the contribution of each stage on the identical 95 corruption–severity environments and scoring samples of the frozen CIFAR-100-C structural audit. The raw source, AdaBN-only, anchor-only, and complete recover–identify–project predictors attain 50.7941%, 55.2427%, 57.9200%, and 61.7177%, respectively. AdaBN-only uses the argmax prediction of the same recalibrated probabilities used to construct the component system. Anchor-only retains the same components, anchors, and fallback rule but removes the pairwise graph term. Thus the complete pipeline improves by $+6.4750$ points over AdaBN-only and by $+3.7977$ points over anchor-only, while its total improvement over the raw source is $+10.9236$ points. Resolved samples

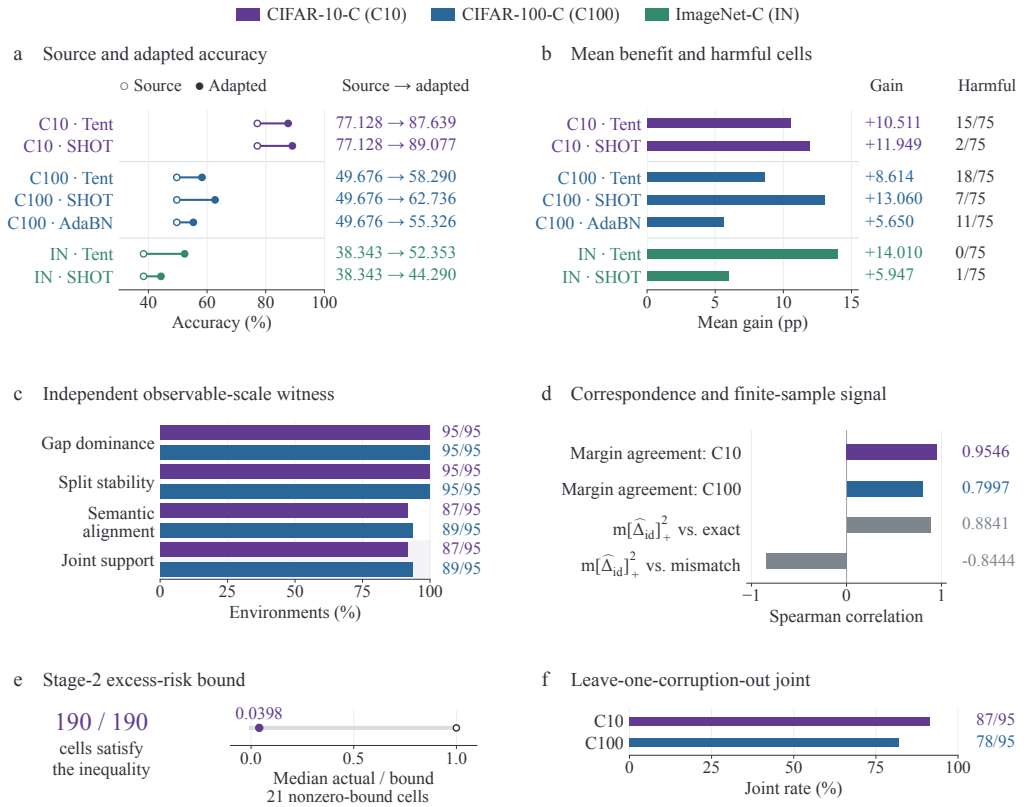

Figure 5: Adaptation outcomes, independent observable-scale evidence, and bound diagnostics. (a,b) Source and adapted accuracy, mean gain in percentage points, and harmful-cell counts on each complete 15×5 corruption grid. (c) The independent source-output merge-gap witness over 95 environments per benchmark; this uses a different component system and is not the A1–A4 joint rate. (d) Post-hoc versus label-free margin correspondence and finite-sample correlations with $\widehat{\Delta}_{\mathrm{id}}^2 m$. (e) Coverage of the Stage-2 excess-risk inequality and the median actual-to-bound ratio among 21 nonzero-bound cells. (f) Corruption-level leave-one-out joint rates. C10, C100, and IN denote CIFAR-10-C, CIFAR-100-C, and ImageNet-C.

receive their recovered component label and unresolved samples retain the fallback prediction of the frozen audit; no additional fitting or target-label information is introduced into this prediction rule. The component oracle uses the same partition, scoring samples, and fallback, with the majority target label assigned to each resolved component. All 79 environments with exact component-label recovery attain their corresponding component-oracle risk. Moreover, all $95/95$ environments satisfy $0 \le \widehat{L}^{(e)}(\widehat{h}) - \widehat{L}_C^{\Gamma,\star,(e)} \le \widehat{M}_{\mathrm{final}}^{(e)}$, where $\widehat{M}_{\mathrm{final}}^{(e)}$ is the final mismatch mass normalized by the full scoring sample in environment e.

For each environment e, define the empirical source-to-resolution gap by $\widehat{G}_s^{(e)} := \widehat{L}^{(e)}(s) - (\widehat{L}_C^{\Gamma,\star,(e)} + \widehat{\zeta}^{(e)} + 2\widehat{\alpha}_{\mathrm{tot}}^{(e)})$, following Section 3.2. Let $\mathcal{E}_{\mathrm{joint}}$ contain the 78 environments simultaneously satisfying the audited A1–A4 diagnostics, and let $\mathcal{E}_+ := \{e \in \mathcal{E}_{\mathrm{joint}} : \widehat{G}_s^{(e)} > 0\}$. This subset contains 13 environments and is defined by the structural diagnostics and gap sign, without selecting on the realized accuracy gain. The recover–identify–project predictor strictly improves over the deployed source in $13/13$ environments in $\mathcal{E}_+$.

The three component statistics reported below are conditional means over this same subset. For any per-environment quantity $q^{(e)}$, write $\overline{q}_+ := |\mathcal{E}_+|^{-1} \sum_{e \in \mathcal{E}_+} q^{(e)}$. Then $\overline{\widehat{L}_C^{\Gamma,\star}}_+ = 0.1827$, $\overline{\widehat{\zeta}}_+ = 0.2451$, and $\overline{\widehat{\alpha}_{\mathrm{tot}}}_+ = 0.1557$. Each constituent risk or mass retains its original full-sample

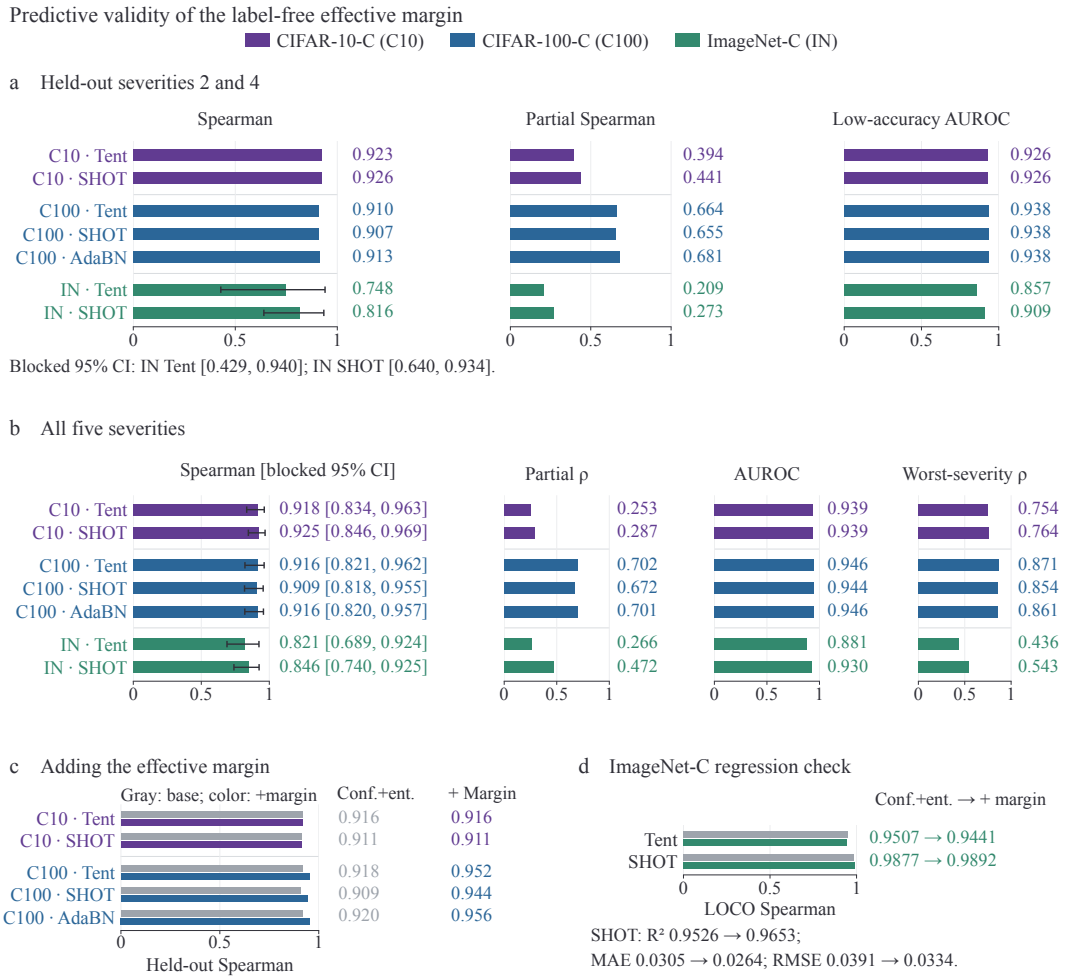

Figure 6: Predictive validity of the label-free effective margin. (a) Held-out severities 2 and 4 following configuration selection on severities 1, 3, and 5. (b) Full-grid rank correlations and within-severity diagnostics. (c) Held-out confidence-plus-entropy prediction with and without the margin. (d) ImageNet-C SHOT leave-one-corruption-out regression, comparing confidence-plus-entropy prediction with and without the effective margin.

denominator within its environment; only the set of environments being averaged changes. These conditional means characterize the jointly admissible, positive-gap regime, while the source and pipeline accuracies above summarize the full 95-environment audit. The $13/13$ result is an empirical, post-hoc check of the source-improvement implication; the finite-sample theoretical guarantee additionally uses the remainder condition $\mathfrak{E}_m < G_s(Q, C)$ stated in Section 3.2.

The source accuracy here is computed on the 19×5 structural-audit population and should not be conflated with the separate 15×5 predictive-study source baseline reported later in this appendix.

B.2 Structural conditions and semantic recovery

We first examine whether the structural conditions underlying the theory are jointly realizable on benchmark corruptions. Across the 95 corruption–severity environments of each dataset, near-purity (A1), stability (A2), and support (A3) hold in 89, 93, and 91 environments on CIFAR-10-C, and in 88, 95, and 86 environments on CIFAR-100-C, respectively. All four conditions hold simultaneously in $86/95$ (90.53%) and $78/95$ (82.11%) environments (Figure 4a). These results show that the complete recoverable-semantic-structure regime arises across a substantial fraction of the audited natural corruption environments on both benchmarks.

Table 3: CIFAR-100-C accuracy decomposition under the frozen structural audit. All variants in the first block use the identical 95 environments and scoring samples. Anchor-only retains the same component system, anchors, and fallback while removing the pairwise graph term.

Quantity	Result
Full audit: 95 *environments*	
Raw source accuracy	50.7941%
AdaBN-only accuracy	55.2427%
Anchor-only accuracy	57.9200%
Recover–identify–project accuracy	**61.7177%**
RIP gain over raw source	**+10.9236** pp
RIP gain over AdaBN-only	**+6.4750** pp
RIP gain over anchor-only	**+3.7977** pp
Joint structural, positive-gap subset: 13 *environments*	
Mean component-oracle risk	0.1827
Mean unresolved mass $\widehat{\zeta}$	0.2451
Mean component impurity $\widehat{\alpha}_{\mathrm{tot}}$	0.1557
RIP $>$ source in $\mathcal{E}_+$	**13/13**

The effective margin also distinguishes successful from unsuccessful semantic identification under the audited configuration. A positive $\widehat{\Delta}_{\mathrm{id}}$ occurs in $90/95$ CIFAR-10-C environments and $79/95$ CIFAR-100-C environments. Across all 190 environments, **these positive-margin indicators agree exactly with exact recovery from random initialization on every environment**. This correspondence supports the operational relevance of the sufficient identifiability condition: in this audit, the balance between semantic support and leakage separates environments in which the true component labeling is recovered from those in which residual errors remain.

Recovery requires substantial correction of the initial labels. Across random, cyclic, and constant initializations, the initial component-mass-weighted mismatch ranges from 0.77223 to 0.87064 on CIFAR-10-C and from 0.49968 to 0.50494 on CIFAR-100-C. All three initializations reach the same reported mean final mismatch of 0.00092 and 0.00157, respectively, and the same exact-recovery counts of $90/95$ and $79/95$ (Figure 4b). Every environment contains at least one accepted label move. Thus, the empirical learner can recover semantic assignments from markedly incorrect starting points, providing algorithmic evidence that the identifiable solutions can be reached under the tested initialization schemes.

B.3 Finite-sample behavior and source orientation

To examine the sample-size dependence predicted by Theorem 3.2, we fix severity 3 and vary the objective sample size over $m \in \{125, 250, 500, 1000, 2500, 5000\}$. Increasing m from 125 to 5000 reduces mean mismatch from 0.00926 to 0 and raises exact recovery from 68.42% to 100% on CIFAR-10-C. On CIFAR-100-C, mismatch decreases from 0.23916 to 0.00099, while exact recovery rises from 0% to 84.21% (Figure 4c–d). Across the evaluated sizes and environments, the positive-margin sample-size score $s_m = m[\widehat{\Delta}_{\mathrm{id}}]_+^2$, where $[z]_+ = \max\{z, 0\}$, has Spearman correlation 0.8841 with exact recovery and -0.8444 with mismatch (Figure 5d). The endpoint improvements and rank associations support the predicted dependence of identification on sample size and effective margin.

The spectral construction also retains the relevant identifiability information in these environments. The certificate of Proposition A.23 is a valid lower bound in all 190 environments and preserves the sign of the effective margin in every case. Its maximum absolute discrepancy from the exact value is 1.11×10^{-10} on CIFAR-10-C and 1.12×10^{-8} on CIFAR-100-C (Figure 4e). These results demonstrate that the computable lower certificate can preserve the empirical positive-margin decision. The near equality here reflects the anchor-dominated graphs; the theoretical guarantee requires a valid lower bound, without assuming such tightness for general graphs.

A controlled intervention isolates the contribution of source orientation. Keeping the component graph fixed, setting $\lambda = 0$ produces no exact recoveries in the 19 severity-3 environments of either benchmark. Introducing anchors with $\lambda = 0.001$ raises exact recovery to $19/19$ on CIFAR-10-C

and $16/19$ on CIFAR-100-C (Figure 4f). Changing only the anchor coefficient therefore restores semantic recovery in most environments. This supports the mechanism that source information resolves the semantic orientation left undetermined by graph structure alone, as required by the theory's distinction between geometric structure and identifiable class labels.

B.4 Independent observable-scale witness

The source-output merge-gap construction of Appendix A provides a complementary test of whether a stable component scale can be selected from observable information. Gap dominance and split stability both hold in all $95/95$ environments of each benchmark. Under this construction's specified semantic-alignment criteria, alignment and joint support hold in $87/95$ (91.58%) CIFAR-10-C environments and $89/95$ (93.68%) CIFAR-100-C environments (Figure 5c). These results support the feasibility of constructing an observable, stable scale that also meets the semantic criterion in most audited environments. The remaining alignment failures show why observable stability and semantic validity must be assessed separately.

B.5 Predictive validity of a label-free effective margin

We next examine whether the effective-margin perspective provides an observable indicator of actual adaptation outcomes. On the complete 15×5 corruption grids, Tent (Wang et al., 2020) and SHOT (Liang et al., 2020) improve CIFAR-10-C accuracy from 77.128% to 87.639% and 89.077%, respectively.On CIFAR-100-C, Tent and SHOT (Liang et al., 2020) improve the 49.676% source accuracy to 58.290% and 62.736%, respectively, while AdaBN (Li et al., 2016) yields a mean gain of $+5.65$ percentage points. On ImageNet-C, Tent and SHOT increase accuracy from 38.343% to 52.353% and 44.290% (Figure 5a–b). These positive mean gains coexist with heterogeneous cell-level outcomes: Tent decreases accuracy in $15/75$ CIFAR-10-C and $18/75$ CIFAR-100-C cells, compared with $2/75$ and $7/75$ for SHOT; AdaBN is harmful in $11/75$ CIFAR-100-C cells. On ImageNet-C, the corresponding counts are $0/75$ for Tent and $1/75$ for SHOT. This variation motivates evaluating whether the margin can rank the accuracy attained after adaptation.

For this purpose, we construct a separate label-free margin $\widehat{\Delta}^{u}_{\mathrm{id}}$ using only source outputs and unlabeled target features at evaluation time. Projected source-model features are clustered with MiniBatchKMeans, mean source posteriors with class-frequency balancing define pseudo component labels, and a shifted-cosine graph with source anchors supplies the margin. Configurations are selected using adaptation outcomes on severities $1, 3, 5$ and then fixed for evaluation on held-out severities $2, 4$. On these held-out severities, Spearman correlations with adapted accuracy are 0.923 and 0.926 for CIFAR-10-C Tent and SHOT, and 0.910, 0.907, and 0.913 for CIFAR-100-C Tent, SHOT, and AdaBN. On ImageNet-C, the correlations are 0.748 for Tent and 0.816 for SHOT, with corruption-blocked 95% intervals $[0.429, 0.940]$ and $[0.640, 0.934]$ (Figure 6a). Low-accuracy detection AUROC is 0.926 on CIFAR-10-C, 0.938 on CIFAR-100-C, and 0.857 and 0.909 for ImageNet-C Tent and SHOT. These results support the margin's practical value as an environment-level indicator of adapted accuracy, including severities excluded from configuration selection.

The association also persists within severity levels. Over all five severities, Spearman correlations range from 0.909 to 0.925 across the CIFAR-C methods and reach 0.821 and 0.846 for ImageNet-C Tent and SHOT. Every reported corruption-blocked interval excludes zero. The weakest within-severity correlations are 0.754 and 0.764 on CIFAR-10-C, 0.854–0.871 on CIFAR-100-C, and 0.436 and 0.543 on ImageNet-C (Figure 6b). Thus, the observed association cannot be attributed solely to differences in average difficulty across severities. These full-grid diagnostics complement the held-out results; they include the configuration-selection severities.

Controlling for mean source confidence by rank residualization leaves positive held-out partial correlations: 0.394–0.441 on CIFAR-10-C, 0.655–0.681 on CIFAR-100-C, and 0.209–0.273 on ImageNet-C. The strongest evidence of complementary predictive information occurs on CIFAR-100-C, where adding margin features to confidence and normalized entropy increases held-out Spearman from 0.918 to 0.952 for Tent, 0.909 to 0.944 for SHOT, and 0.920 to 0.956 for AdaBN.

ImageNet-C leave-one-corruption-out regression provides complementary evidence. For SHOT, adding margin features improves Spearman from 0.9877 to 0.9892 and R^2 from 0.9526 to 0.9653, while reducing MAE from 0.0305 to 0.0264 and RMSE from 0.0391 to 0.0334 (Figure 6d). This pro-

vides evidence that the effective margin contains predictive information complementary to confidence and entropy.

B.6 Correspondence and bound diagnostics

The post-hoc structural margin and the label-free predictive margin have Spearman correlations of 0.9546 on CIFAR-10-C and 0.7997 on CIFAR-100-C (Figure 5d). Their similar environment orderings connect the audited identifiability analysis with the observable predictive construction.

The empirical Stage-2 inequality relating excess risk to component mismatch holds in all $190/190$ audited environments (Figure 5e). This confirms the predicted ordering between component-level semantic identification error and downstream excess risk on the audited environments.

Finally, we assess sensitivity to the label-informed outer configuration-selection step using corruption-level leave-one-out selection. In each fold, all five severities of one corruption family are excluded from selection; the same exploratory selection procedure is applied to the remaining families, and the selected configuration is evaluated on the held-out family. Pooling the held-out environments yields joint structural satisfaction in $87/95$ CIFAR-10-C and $78/95$ CIFAR-100-C environments, compared with $86/95$ and $78/95$ for the single reported dataset-level configurations. Because the candidate configurations and selection procedure were developed on these benchmarks, this is an exploratory robustness check, not untouched validation or label-free model selection.

C Proofs for Finite-Width Entropy Dynamics

C.1 Preliminaries and proof structure

This appendix proves the finite-width entropy-minimization results stated in Section 4. We first restate the binary finite-pool setting and the source-oriented component structure used by the analysis. We then derive the exact score update for a finite-width two-layer ReLU network with moving gates, relate component-level semantic support to a pointwise corrective field, and control the change of this field over an early adaptation window. These ingredients yield the full-batch guarantee in Theorem 4.1 and its minibatch extension. The final part of the appendix gives one sufficient finite-width construction under which a positive corrective field is attainable; this construction is not an assumption of the main theorem.

Binary component specialization. As in Section 4, let $\mathcal{X}_T = \{x_1, \dots, x_N\}$ be the finite target pool and let

$$\mu := \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i} \tag{41}$$

be its empirical probability measure. We assume $\|x\|_2 \le 1$ for every $x \in \mathcal{X}_T$. The mechanism analysis considers the binary, pure-component specialization of Section 3: before target adaptation, the pool is partitioned into nonempty semantic components $C_1, \dots, C_M$, with masses $m_u := \mu(C_u) > 0$. Each component is assigned an orientation $\epsilon_u \in \{-1, +1\}$, and

$$y^\star(x) = \epsilon_u, \qquad x \in C_u. \tag{42}$$

We write $\epsilon(x) = \epsilon_u$ for $x \in C_u$. This is the $\alpha_u = 0$ binary case of the semantic-component model used in the PAC analysis. The component system and the orientations are analysis objects: neither is supplied to the entropy update.

For a score $s_t(x) \in \mathbb{R}$, define the signed margin and the empirical target risk by

$$z_t(x) := \epsilon(x) s_t(x), \qquad \mathcal{R}_t := \mu\big(\{x \in \mathcal{X}_T : z_t(x) \le 0\}\big). \tag{43}$$

Thus a point is correct exactly when $z_t(x) > 0$; ties are counted as errors.

Network and adapted parameters. Let $D = d_{\text{in}} + 1$, set $\widetilde{x} = (x, 1) \in \mathbb{R}^D$, and write $\widetilde{w}_j = (w_j, b_j)$. The model is the width-m_h two-layer ReLU network

$$s_t(x) = \sum_{j=1}^{m_h} a_j [\langle \widetilde{w}_{j,t}, \widetilde{x} \rangle]_+, \tag{44}$$

with fixed output coefficients $\{a_j\}_{j=1}^{m_h}$. Target adaptation updates only the hidden parameters $\vartheta_t = (\widetilde{w}_{1,t}, \dots, \widetilde{w}_{m_h,t})$. Define

$$u_{j,t}(x) := \langle \widetilde{w}_{j,t}, \widetilde{x} \rangle, \qquad g_{j,t}(x) := \mathbf{1}\{u_{j,t}(x) > 0\}, \qquad A_2 := \sum_{j=1}^{m_h} a_j^2. \tag{45}$$

Because $\|x\|_2, \|x'\|_2 \le 1$, $0 \le \langle \widetilde{x}, \widetilde{x}' \rangle = 1 + x^\top x' \le 2$.

Source-oriented cores. For each component C_u, let $G_u \subseteq C_u$ be a *source-oriented core* and let $B_u := C_u \setminus G_u$ be its *uncertain shell*. We use the unions

$$G_\epsilon := \bigcup_{u:\epsilon_u=\epsilon} G_u, \qquad B := \bigcup_{u=1}^{M} B_u. \tag{46}$$

The only property required by the component-to-field argument is that the pretrained score has the correct sign on G_u. Lemma C.1 below gives a sufficient score-level certificate from source anchors and an input-Lipschitz bound.

Proof structure. Section C.2 records the properties of the entropy force and proves the source-core certificate. Section C.3 derives the exact finite-width score dynamics, including the ReLU switching remainder. Section C.4 proves the component-to-pointwise support–leakage bridge used in the main text. Section C.5 controls gate, kernel, score, and field movement from quantities fixed before adaptation. Section C.6 then proves full-batch and minibatch risk descent. Finally, Section C.8 gives a sufficient initialization and source-movement regime under which a positive pre-adaptation corrective field exists at finite width.

C.2 Entropy force and source-oriented cores

Let $p(s) = \sigma(s)$ and let $h(s) = -p(s)\log p(s) - (1-p(s))\log(1-p(s))$ denote binary predictive entropy. Its negative derivative with respect to the score is

$$\psi(s) := -h'(s) = s\,p(s)(1-p(s)) = \frac{s}{2 + 2\cosh s}. \tag{47}$$

Consequently,

$$\operatorname{sign}\psi(s) = \operatorname{sign} s, \qquad \psi(-s) = -\psi(s). \tag{48}$$

This is the sense in which entropy has no pointwise correctness signal: an isolated negative score produces a negative entropy force and is therefore pushed further in its current direction.

For later bounds we record two elementary properties of ψ. For $s > 0$,

$$\psi'(s) = \frac{1 + \cosh s - s\sinh s}{2(1+\cosh s)^2},$$

and the numerator has derivative $-s\cosh s < 0$. Hence ψ increases and then decreases on $(0,\infty)$, so for $0 < a \le b$,

$$\min_{z\in[a,b]} \psi(z) = \min\{\psi(a), \psi(b)\}. \tag{49}$$

Moreover,

$$\overline{\psi} := \sup_{s\in\mathbb{R}} |\psi(s)| < \infty, \qquad \|\psi'\|_\infty \le \frac{1}{4}. \tag{50}$$

The second inequality follows directly by differentiating $\psi(s) = s/(2+2\cosh s)$.

We next give the source-anchor certificate used in the main text. For each component, let $A_u \subseteq C_u$ be a set of source anchors whose pretrained score has the correct orientation with margin $\gamma_a > 0$:

$$\epsilon_u s_0(z) \ge \gamma_a \qquad \forall z \in A_u. \tag{51}$$

This is the binary score-level counterpart of the source-oriented anchor information in Section 3.1. Define

$$L_0 := \sum_{j=1}^{m_h} |a_j|\,\|w_{j,0}\|_2. \tag{52}$$

Since ReLU is 1-Lipschitz, L_0 is an input-Lipschitz constant for s_0. Fix $r_a > 0$ and take $G_u := \{x \in C_u : \operatorname{dist}(x, A_u) \le r_a\}$. Let $\gamma_c := \gamma_a - L_0 r_a$.

Lemma C.1 (Source anchors certify core orientation). *If $\gamma_c > 0$, then*

$$\epsilon_u s_0(x) \geq \gamma_c > 0 \qquad \forall x \in G_u. \tag{53}$$

If, in addition, $|s_0(x)| \leq S$ on $\mathcal{X}_T \cup \bigcup_u A_u$, then

$$\epsilon_u \psi(s_0(x)) \geq \psi_{\min}(0) := \min\{\psi(\gamma_c), \psi(S)\} > 0 \qquad \forall x \in G_u. \tag{54}$$

Proof. For any x, z,

$$|s_0(x) - s_0(z)| \leq \sum_j |a_j| \, |[w_{j,0}^\top x + b_{j,0}]_+ - [w_{j,0}^\top z + b_{j,0}]_+| \leq L_0 \|x - z\|_2.$$

If $x \in G_u$, choose $z \in A_u$ with $\|x - z\|_2 \leq r_a$. Then $\epsilon_u s_0(x) \geq \gamma_a - L_0 r_a = \gamma_c$. Under the score bound, $\epsilon_u s_0(x) \in [\gamma_c, S]$. Oddness of ψ and (49) give the force bound. □

C.3 Exact finite-width score dynamics

At target step t, let $X_{t,1}, \ldots, X_{t,B} \overset{\text{iid}}{\sim} \mu$ be a minibatch sampled with replacement from the finite target pool. Gradient descent on predictive entropy updates the hidden parameters as

$$\widetilde{w}_{j,t+1} = \widetilde{w}_{j,t} + \eta a_j \frac{1}{B} \sum_{b=1}^{B} \psi(s_t(X_{t,b})) g_{j,t}(X_{t,b}) \widetilde{X}_{t,b}. \tag{55}$$

The full-batch dynamics used in Theorem 4.1 are obtained by replacing the minibatch average with the exact empirical average under μ.

For a receiver $x' \in \mathcal{X}_T$, define the induced change of the jth preactivation by

$$\delta_{j,t}(x') := \langle \widetilde{w}_{j,t+1} - \widetilde{w}_{j,t}, \widetilde{x}' \rangle \tag{56}$$

and define the exact ReLU switching residual

$$r_{j,t+1}(x') := [u_{j,t}(x') + \delta_{j,t}(x')]_+ - [u_{j,t}(x')]_+ - g_{j,t}(x')\delta_{j,t}(x'). \tag{57}$$

Set $R_{t+1}(x') := \sum_j a_j r_{j,t+1}(x')$. The scalar residual vanishes if the gate at x' does not switch, and in all cases

$$|r_{j,t+1}(x')| \leq |\delta_{j,t}(x')| \mathbf{1}\{g_{j,t+1}(x') \neq g_{j,t}(x')\}. \tag{58}$$

The hidden-parameter coupling kernel is the exact Jacobian inner product

$$\begin{aligned} K_t(x, x') &:= \langle \nabla_\vartheta s_t(x), \nabla_\vartheta s_t(x') \rangle \\ &= (1 + x^\top x') \sum_{j=1}^{m_h} a_j^2 g_{j,t}(x) g_{j,t}(x'). \end{aligned} \tag{59}$$

Hence $K_t(x, x') \geq 0$ pointwise and $K_t(x, x') \leq 2A_2$. Define the influence field

$$F_t(x') := \int K_t(x, x') \psi(s_t(x)) \, d\mu(x) \tag{60}$$

and the centered minibatch fluctuation

$$\xi_{t+1}(x') := \frac{1}{B} \sum_{b=1}^{B} K_t(X_{t,b}, x') \psi(s_t(X_{t,b})) - F_t(x'). \tag{61}$$

Proposition C.2 (Exact finite-width score update). *For every receiver $x' \in \mathcal{X}_T$,*

$$s_{t+1}(x') - s_t(x') = \eta F_t(x') + \eta \xi_{t+1}(x') + R_{t+1}(x'), \tag{62}$$

and $\mathbb{E}[\xi_{t+1}(x') \mid \mathcal{F}_t] = 0$, where $\mathcal{F}_t$ contains the initial model and all target-time randomness revealed before the step-t minibatch is drawn.

Proof. By definition of $r_{j,t+1}$,

$$[u_{j,t} + \delta_{j,t}]_+ - [u_{j,t}]_+ = g_{j,t}\delta_{j,t} + r_{j,t+1}.$$

Therefore

$$s_{t+1}(x') - s_t(x') = \sum_j a_j g_{j,t}(x')\delta_{j,t}(x') + R_{t+1}(x').$$

Substituting (55) into $\delta_{j,t}(x')$ gives

$$\begin{aligned}\sum_j a_j g_{j,t}(x')\delta_{j,t}(x') &= \eta\frac{1}{B}\sum_{b=1}^{B}\psi(s_t(X_{t,b}))\langle\widetilde{X}_{t,b},\widetilde{x}'\rangle\sum_j a_j^2 g_{j,t}(X_{t,b})g_{j,t}(x') \\ &= \eta\frac{1}{B}\sum_{b=1}^{B}K_t(X_{t,b},x')\psi(s_t(X_{t,b})) \\ &= \eta F_t(x') + \eta\xi_{t+1}(x').\end{aligned}$$

This proves (62). Conditional on $\mathcal{F}_t$, the integrand is fixed and the minibatch samples are iid from μ, which gives the conditional mean-zero property. □

Proposition C.2 is an identity for the actual finite network. In particular, no frozen-gate linearization or infinite-width limit is used; all changes caused by crossing ReLU boundaries are contained in R_{t+1}.

C.4 From semantic support to a pointwise corrective field

We now connect the source-oriented component structure to the initial influence field. For a receiver component C_u, define its average signed field by

$$\bar{F}_u := \frac{1}{m_u}\int_{C_u}\epsilon_u F_0(x')\,d\mu(x'). \tag{63}$$

The same-orientation source cores contribute with the correct sign, whereas opposite-orientation cores and all uncertain shells are bounded conservatively. Accordingly, define

$$S_u^J := \frac{1}{m_u}\int_{C_u}\int_{G_{\epsilon_u}} K_0(x,x')|\psi(s_0(x))|\,d\mu(x)d\mu(x'), \tag{64}$$

$$L_u^J := \frac{1}{m_u}\int_{C_u}\int_{G_{-\epsilon_u}\cup B} K_0(x,x')|\psi(s_0(x))|\,d\mu(x)d\mu(x'). \tag{65}$$

Let

$$\Delta_J := \min_u(S_u^J - L_u^J), \qquad \omega_J := \max_u \sup_{x'\in C_u}|\epsilon_u F_0(x') - \bar{F}_u|. \tag{66}$$

Here Δ_J is a component-average support–leakage margin, while ω_J measures the within-component variation that must be paid when passing from an average statement to a pointwise one.

Proposition C.3 (Component-to-pointwise semantic propagation)**.** *If the cores satisfy the source-orientation property of Lemma C.1, then*

$$\bar{F}_u \geq S_u^J - L_u^J \qquad \forall u, \tag{67}$$

and hence

$$d_0 := \min_{x'\in\mathcal{X}_T}\epsilon(x')F_0(x') \geq \Delta_J - \omega_J. \tag{68}$$

Proof. Fix u. If $x \in G_{\epsilon_u}$, then the source-oriented core property and (48) imply $\epsilon_u\psi(s_0(x)) = |\psi(s_0(x))|$. For $x \in G_{-\epsilon_u}\cup B$, we only use $\epsilon_u\psi(s_0(x)) \geq -|\psi(s_0(x))|$. Since $K_0 \geq 0$, splitting the sender integral in (60) over these two sets and averaging over $x' \in C_u$ proves (67). For any receiver $x' \in C_u$,

$$\epsilon_u F_0(x') \geq \bar{F}_u - |\epsilon_u F_0(x') - \bar{F}_u| \geq (S_u^J - L_u^J) - \omega_J \geq \Delta_J - \omega_J.$$

Taking the minimum over all receivers gives (68). □

Relation to the PAC identification margin. The quantity Δ_J is the optimizer-specific analogue of the static support–leakage margin Δ_{id} in Section 3.1. The two quantities are not asserted to be numerically equal: Δ_{id} is defined from a source-induced component graph, whereas Δ_J is formed from the actual parameter-gradient coupling kernel K_0 and the entropy-force magnitudes. When the component graph is induced by K_0 and source orientation is represented through the core forces, the two decompositions have the same support-versus-leakage form. Proposition C.3 is the additional step required by the dynamics, because score correction needs a pointwise field rather than a component-average objective gap.

C.5 Stability over an early finite-width trajectory

A positive initial field need not remain positive if the hidden parameters move enough to alter many ReLU gates or substantially change the scores. We therefore control the early trajectory by quantities fixed at $t = 0$. Fix a movement budget $\Lambda > 0$ and define the initial gate-boundary energy

$$\Gamma_0(\Lambda) := \max_{x \in \mathcal{X}_T} \sum_{j=1}^{m_h} a_j^2 \mathbf{1}\{|u_{j,0}(x)| \le 2\Lambda |a_j| \overline{\psi}\}. \tag{69}$$

Thus $\Gamma_0(\Lambda)$ is the maximum a_j^2-weighted mass of hidden units whose initial preactivation lies close enough to zero to change sign within the prescribed movement budget. Also set

$$\sigma_{\mathrm{drift}} := 2\Lambda A_2 \overline{\psi}. \tag{70}$$

Lemma C.4 (Finite-movement bounds)**.** *If $\eta(T+1) \le \Lambda$, then for every $t \le T$ and all target points,*

$$|u_{j,t}(x) - u_{j,0}(x)| \le 2\Lambda |a_j| \overline{\psi}, \tag{71}$$

$$\sum_{j=1}^{m_h} a_j^2 |g_{j,t}(x) - g_{j,0}(x)| \le \Gamma_0(\Lambda), \tag{72}$$

$$|K_t(x,x') - K_0(x,x')| \le 4\Gamma_0(\Lambda), \tag{73}$$

$$|s_t(x) - s_0(x)| \le \sigma_{\mathrm{drift}}, \tag{74}$$

$$|R_{t+1}(x')| \le 2\eta \overline{\psi} \Gamma_0(\Lambda). \tag{75}$$

Proof. Equation (55) and $|\langle \widetilde{X}, \widetilde{x} \rangle| \le 2$ give $|\delta_{j,t}(x)| \le 2\eta |a_j| \overline{\psi}$. Summing over the trajectory yields (71). A gate can differ from its initial value only if its initial preactivation lies within the maximal possible displacement of zero, which gives (72) by the definition of $\Gamma_0(\Lambda)$.

For binary gates, $|ab - cd| \le |a - c| + |b - d|$. Together with $|1 + x^\top x'| \le 2$, this yields

$$|K_t(x,x') - K_0(x,x')| \le 2 \sum_j a_j^2 \big(|g_{j,t}(x) - g_{j,0}(x)| + |g_{j,t}(x') - g_{j,0}(x')| \big) \le 4\Gamma_0(\Lambda).$$

ReLU Lipschitzness gives $|s_{t+1}(x) - s_t(x)| \le 2\eta A_2 \overline{\psi}$, which sums to (74). Finally, (58) and the fact that every switching gate belongs to the initial danger band imply

$$|R_{t+1}(x')| \le \sum_j |a_j| \, |\delta_{j,t}(x')| \mathbf{1}\{j \text{ switches at } x'\} \le 2\eta \overline{\psi} \Gamma_0(\Lambda).$$

□

Lemma C.5 (Stability of the influence field)**.** *Under the same movement budget,*

$$|F_t(x') - F_0(x')| \le 4\overline{\psi} \Gamma_0(\Lambda) + \frac{A_2}{2} \sigma_{\mathrm{drift}} \qquad \forall t \le T, \ \forall x' \in \mathcal{X}_T. \tag{76}$$

Proof. Add and subtract $\int K_0(x,x') \psi(s_t(x)) \, d\mu(x)$ to obtain

$$\begin{aligned} |F_t(x') - F_0(x')| &\le \int |K_t - K_0| \, |\psi(s_t)| \, d\mu + \int K_0 |\psi(s_t) - \psi(s_0)| \, d\mu \\ &\le 4\overline{\psi} \Gamma_0(\Lambda) + 2A_2 \|\psi'\|_\infty \|s_t - s_0\|_\infty. \end{aligned}$$

Apply (50) and Lemma C.4. □

For the support–leakage certificate of Proposition C.3, define

$$\underline{d}_0 := \Delta_J - \omega_J. \tag{77}$$

Combining the field drift and the exact ReLU remainder motivates the pre-adaptation corrective margin

$$\Delta_{\mathrm{corr}}(\Lambda) := \underline{d}_0 - 6\overline{\psi}\Gamma_0(\Lambda) - \frac{A_2}{2}\sigma_{\mathrm{drift}}. \tag{78}$$

All terms in (78) are fixed before target adaptation once Λ is chosen. More generally, the arguments below remain valid if $\underline{d}_0$ is replaced by any certified lower bound on d_0; the support–leakage bridge is the particular certificate used in the main text.

C.6 Full-batch and minibatch risk descent

Theorem C.6 (Full-batch finite-width correction)**.** *Suppose $\Delta_{\mathrm{corr}}(\Lambda) > 0$ and full-batch entropy gradient descent satisfies $\eta(T+1) \le \Lambda$. Then, for every receiver $x' \in \mathcal{X}_T$ and every $t < T$,*

$$\epsilon(x')[s_{t+1}(x') - s_t(x')] \ge \eta\Delta_{\mathrm{corr}}(\Lambda) > 0. \tag{79}$$

Consequently,

$$z_T(x') \ge z_0(x') + \eta T\Delta_{\mathrm{corr}}(\Lambda) \qquad \forall x' \in \mathcal{X}_T. \tag{80}$$

In particular, no initially correct point becomes incorrect, and every initial error satisfying $-\eta T\Delta_{\mathrm{corr}}(\Lambda) < z_0(x') \le 0$ is corrected. Hence

$$\mathcal{R}_T \le \mathcal{R}_0 - \mu(\{x : -\eta T\Delta_{\mathrm{corr}}(\Lambda) < z_0(x) \le 0\}). \tag{81}$$

Proof. For full-batch updates, $\xi_{t+1} = 0$. Proposition C.3 and Lemma C.5 give

$$\epsilon(x')F_t(x') \ge \underline{d}_0 - 4\overline{\psi}\Gamma_0(\Lambda) - \frac{A_2}{2}\sigma_{\mathrm{drift}}.$$

Using the exact score update (62) and $\epsilon(x')R_{t+1}(x') \ge -|R_{t+1}(x')|$, the remainder bound in Lemma C.4 subtracts a further $2\eta\overline{\psi}\Gamma_0(\Lambda)$. This proves (79). Summing over $t = 0, \dots, T-1$ gives (80). Since every signed margin is nondecreasing by a strictly positive amount, initially correct points cannot cross the decision boundary in the wrong direction, while every error shallower than the accumulated lift crosses to positive signed margin. The risk bound follows. □

Theorem 4.1 is Theorem C.6 with the initial-field certificate $\underline{d}_0 = \Delta_J - \omega_J$ from Proposition C.3.

Minibatch noise. For a receiver x', define the martingale prefix

$$M_t(x') := \eta \sum_{q=0}^{t-1} \xi_{q+1}(x'), \qquad M_0(x') = 0. \tag{82}$$

Since $|K_t\psi(s_t)| \le 2A_2\overline{\psi}$,

$$|\eta\xi_{t+1}(x')| \le 4\eta A_2\overline{\psi}, \qquad \mathrm{Var}(\eta\xi_{t+1}(x') \mid \mathcal{F}_t) \le \frac{4\eta^2 A_2^2\overline{\psi}^2}{B}. \tag{83}$$

Let

$$\ell_N := \log\frac{2N}{\delta_T}, \qquad \mathfrak{M}_T(\delta_T) := 2\eta A_2\overline{\psi}\left[\sqrt{\frac{2T\ell_N}{B}} + \frac{4}{3}\ell_N\right]. \tag{84}$$

Freedman's maximal inequality, followed by a union bound over the N fixed receivers, yields

$$\Pr\left[\max_{t\le T}|M_t(x')| \le \mathfrak{M}_T(\delta_T)\ \forall x' \in \mathcal{X}_T \,\middle|\, \theta_0\right] \ge 1 - \delta_T. \tag{85}$$

Theorem C.7 (Minibatch finite-width correction)**.** *Under the same structural and movement assumptions as Theorem C.6, with probability at least $1 - \delta_T$,*

$$z_T(x') \ge z_0(x') + \eta T\Delta_{\mathrm{corr}}(\Lambda) - \mathfrak{M}_T(\delta_T) \qquad \forall x' \in \mathcal{X}_T. \tag{86}$$

If $\eta T\Delta_{\mathrm{corr}}(\Lambda) > \mathfrak{M}_T(\delta_T)$, no initially correct point becomes incorrect. Every initial error of depth

$$-z_0(x') < \eta T\Delta_{\mathrm{corr}}(\Lambda) - \mathfrak{M}_T(\delta_T) \tag{87}$$

is corrected. Therefore $\mathcal{R}_T < \mathcal{R}_0$ whenever this set of shallow initial errors has positive μ-mass.

Proof. The deterministic movement bounds of Lemma C.4 hold for minibatch updates because they use only the uniform bound on the minibatch average. Hence Proposition C.2 gives

$$z_{t+1}(x') - z_t(x') \geq \eta \Delta_{\mathrm{corr}}(\Lambda) + \eta \epsilon(x') \xi_{t+1}(x').$$

Summing over t and applying (85) proves (86). Multiplication by $\epsilon(x') \in \{-1, +1\}$ does not change the absolute martingale bound. The claims about sign preservation and correction then follow as in the full-batch case. □

C.7 Trajectory-wise exact certificate

The pre-adaptation theorem above controls an entire early window using only initial quantities and the movement budget. The exact identity also gives a sharper ex post sufficient condition for a realized full-batch trajectory. Define

$$d_t := \min_{x' \in \mathcal{X}_T} \epsilon(x') F_t(x'), \qquad r_{t+1} := \min_{x' \in \mathcal{X}_T} \frac{\epsilon(x') R_{t+1}(x')}{\eta}. \tag{88}$$

Proposition C.8 (Realized finite-trajectory certificate)**.** *For full-batch adaptation, if $d_t + r_{t+1} > 0$ for every $t = 0, \ldots, T-1$, then every signed margin increases at every step and*

$$z_T(x') \geq z_0(x') + \eta \sum_{t=0}^{T-1} (d_t + r_{t+1}) \qquad \forall x' \in \mathcal{X}_T. \tag{89}$$

Thus no new error is created, and every initial error shallower than the accumulated lift is corrected.

Proof. In the full-batch case, Proposition C.2 gives

$$\epsilon(x')[s_{t+1}(x') - s_t(x')] = \eta \epsilon(x') F_t(x') + \epsilon(x') R_{t+1}(x') \geq \eta (d_t + r_{t+1}).$$

Summing the inequality over t proves the claim. □

Unlike Theorem C.6, this proposition depends on quantities realized after adaptation begins. It is therefore a trajectory certificate rather than a pre-adaptation condition.

C.8 A sufficient finite-width route to the corrective regime

The preceding results are conditional on a positive initial corrective field. We now give one sufficient construction showing that such a regime is attainable at finite width without assuming the success of the subsequent entropy trajectory. The construction has three ingredients: isotropic random hidden initialization, bounded source-training movement, and a source-trained model that satisfies the core-orientation certificate of Lemma C.1. The geometric conditions introduced below are used only for this sufficient construction and are not assumptions of Theorem C.6.

C.8.1 Random initialization and bounded source movement

At the beginning of source training, let

$$\widetilde{w}_{j,-} = \frac{\sigma_w}{\sqrt{m_h}} G_j, \qquad G_j \overset{\text{iid}}{\sim} \mathcal{N}(0, I_D), \tag{90}$$

and let the fixed output head be

$$a_j = \xi_j \sqrt{\frac{A_2}{m_h}}, \qquad \xi_j \overset{\text{iid}}{\sim} \mathrm{Unif}\{-1, +1\}. \tag{91}$$

The output head is held fixed during both source training and target adaptation. Suppose the derivative of the source loss with respect to the network score is bounded in magnitude by L_S, and let $H_S := \sum_{r < T_S} \eta_{S,r}$ be the cumulative source step size. Since $\|\widetilde{x}\|_2 \leq \sqrt{2}$, every source-training path satisfies

$$\left\| \widetilde{w}_{j,0} - \frac{\sigma_w}{\sqrt{m_h}} G_j \right\|_2 \leq \frac{\sigma_w \tau_S}{\sqrt{m_h}}, \qquad \tau_S := \frac{\sqrt{A_2} L_S \sqrt{2} H_S}{\sigma_w}. \tag{92}$$

Indeed, each source step moves hidden unit j by at most $\eta_{S,r}\sqrt{A_2/m_h}L_S\sqrt{2}$, and (92) follows by summation.

For $r \in [-1,1]$, let

$$q(r) := \Pr(Z_1 > 0, Z_2 > 0) = \frac{1}{4} + \frac{1}{2\pi}\arcsin r, \tag{93}$$

where (Z_1, Z_2) is a standard Gaussian pair with correlation r. For $\delta_S \in (0,1)$, define

$$v_m(\delta_S) := C_{\mathrm{vc}}\sqrt{\frac{D\log(2em_h) + \log(16/\delta_S)}{m_h}} \tag{94}$$

and

$$b_m(t) := \min\left\{1, \sqrt{\frac{2}{\pi}}t + v_m(\delta_S)\right\}. \tag{95}$$

A standard VC uniform-convergence event for Gaussian halfspaces and slabs has probability at least $1-\delta_S$ and simultaneously controls the empirical double-halfspace probabilities by $v_m(\delta_S)$ and the fraction of Gaussian directions satisfying $|\langle G_j, v\rangle| \le t$ by $b_m(t)$, uniformly over unit vectors v.

Lemma C.9 (Finite-width coupling concentration after source training)**.** *On the above VC event, the source-trained gates satisfy, uniformly over $x, x' \in \mathcal{X}_T$,*

$$\left|\frac{1}{A_2}\sum_{j=1}^{m_h} a_j^2 g_{j,0}(x)g_{j,0}(x') - q(\langle\widehat{x},\widehat{x}'\rangle)\right| \le \varepsilon_{\mathrm{gate}}, \tag{96}$$

where

$$\varepsilon_{\mathrm{gate}} := v_m(\delta_S) + 2b_m(\tau_S), \qquad \widehat{x} := \frac{\widetilde{x}}{\|\widetilde{x}\|_2}. \tag{97}$$

Proof. At initialization, $\mathbf{1}\{\langle G_j, \widehat{x}\rangle > 0\}\mathbf{1}\{\langle G_j, \widehat{x}'\rangle > 0\}$ is a Gaussian double-halfspace indicator. The VC event therefore bounds its empirical average by $q(\langle\widehat{x},\widehat{x}'\rangle)$ up to v_m. If source training changes the gate at x, then the initial Gaussian projection must lie within the total source displacement of the ReLU boundary. By (92), this implies $|\langle G_j, \widehat{x}\rangle| \le \tau_S$, whose empirical frequency is at most $b_m(\tau_S)$. Since for binary variables $|ab - cd| \le |a-c| + |b-d|$, the double-gate overlap can change by at most two such slab masses. Finally, $a_j^2 = A_2/m_h$, which gives (96). □

The same event controls the target-time gate-boundary energy.

Lemma C.10 (Gate-boundary energy under random finite width)**.** *Under the initialization and source-movement conditions above, for every $\Lambda > 0$,*

$$\Gamma_0(\Lambda) \le A_2\, b_m\left(\tau_S + \frac{2\Lambda\sqrt{A_2\overline{\psi}}}{\sigma_w}\right) \tag{98}$$

on the same VC event.

Proof. If unit j lies in the target danger band at x, then $|u_{j,0}(x)| \le 2\Lambda|a_j|\overline{\psi}$. Write $\widetilde{w}_{j,0} = (\sigma_w/\sqrt{m_h})G_j + \Delta\widetilde{w}_j$ and use (92). Because $|a_j| = \sqrt{A_2/m_h}$ and $\|\widetilde{x}\|_2 \ge 1$, membership in the danger band implies

$$|\langle G_j, \widehat{x}\rangle| \le \tau_S + \frac{2\Lambda\sqrt{A_2\overline{\psi}}}{\sigma_w}.$$

The slab bound b_m controls the fraction of such units uniformly over x. Multiplying by $a_j^2 = A_2/m_h$ and summing gives the result. □

C.8.2 Orientation separation and a positive initial field

We next impose a sufficient geometric separation between the two semantic orientations. Define

$$C_\epsilon := \bigcup_{u:\epsilon_u=\epsilon} C_u, \qquad m_\epsilon := \mu(C_\epsilon), \qquad \epsilon \in \{-1,+1\}. \tag{99}$$

Assume $m_+, m_- > 0$. Let $c_+, c_- \in \mathbb{S}^{D-1}$ and let α, Δ satisfy

$$\angle(\widehat{x}, c_\epsilon) \le \alpha \quad \forall x \in C_\epsilon, \qquad \angle(c_+, c_-) \ge \Delta, \qquad 0 \le 2\alpha < \Delta - 2\alpha \le \frac{\pi}{2}. \tag{100}$$

Then same-orientation pairs satisfy $\langle \widehat{x}, \widehat{x}' \rangle \ge \rho_{\text{in}} := \cos(2\alpha)$, while opposite-orientation pairs satisfy $0 \le \langle \widehat{x}, \widehat{x}' \rangle \le \rho_{\text{out}} := \cos(\Delta - 2\alpha)$. Define

$$k_{\text{in}} := \rho_{\text{in}} q(\rho_{\text{in}}), \qquad k_{\text{out}} := 2\rho_{\text{out}} q(\rho_{\text{out}}). \tag{101}$$

Assume further that

$$\mu(G_\epsilon) \ge p_\epsilon > 0 \quad (\epsilon \in \{-1,+1\}), \qquad \mu(B) \le \beta_{\text{sh}}. \tag{102}$$

Let $\psi_{\min}(0) > 0$ be the core-force floor from Lemma C.1. Set

$$\kappa_{\text{in}}^{\text{fw}} := A_2(k_{\text{in}} - 2\varepsilon_{\text{gate}}), \tag{103}$$

$$\mathcal{L}_\epsilon^{\text{fw}} := A_2\left[2\beta_{\text{sh}} + (k_{\text{out}} + 2\varepsilon_{\text{gate}}) m_{-\epsilon}\right], \tag{104}$$

$$d_{\text{prim}} := \min_{\epsilon \in \{-1,+1\}} \left\{ \psi_{\min}(0) p_\epsilon \kappa_{\text{in}}^{\text{fw}} - \overline{\psi} \mathcal{L}_\epsilon^{\text{fw}} \right\}. \tag{105}$$

Proposition C.11 (Sufficient finite-width lower bound on the initial field)**.** *On the high-probability coupling-concentration event, if $\kappa_{\text{in}}^{\text{fw}} > 0$, then every receiver satisfies*

$$\epsilon(x') F_0(x') \ge d_{\text{prim}}, \qquad \textit{and hence} \qquad d_0 \ge d_{\text{prim}}. \tag{106}$$

If, in addition,

$$d_{\text{prim}} > 6\overline{\psi}\Gamma_0(\Lambda) + \frac{A_2}{2}\sigma_{\text{drift}}, \tag{107}$$

then the full-batch conclusion of Theorem C.6 holds with d_{prim} used as the certified lower bound on d_0.

Proof. Consider first a same-orientation pair. Writing $r = \langle \widehat{x}, \widehat{x}' \rangle$, we have $r \ge \rho_{\text{in}} \ge 0$ and $1 + x^\top x' = \|\widetilde{x}\|_2 \|\widetilde{x}'\|_2 r$ with $1 \le \|\widetilde{x}\|_2, \|\widetilde{x}'\|_2 \le \sqrt{2}$. Since $r \mapsto rq(r)$ is increasing on $[0,1]$, Lemma C.9 gives

$$K_0(x, x') \ge A_2[k_{\text{in}} - 2\varepsilon_{\text{gate}}] = \kappa_{\text{in}}^{\text{fw}}.$$

For opposite orientations, $0 \le r \le \rho_{\text{out}}$, hence

$$K_0(x, x') \le A_2[k_{\text{out}} + 2\varepsilon_{\text{gate}}].$$

For arbitrary shell senders we use the global bound $K_0 \le 2A_2$. Therefore, for a receiver of orientation ϵ, same-orientation cores contribute at least $\psi_{\min}(0) p_\epsilon \kappa_{\text{in}}^{\text{fw}}$, opposite-orientation cores contribute no less than $-\overline{\psi} A_2 (k_{\text{out}} + 2\varepsilon_{\text{gate}}) m_{-\epsilon}$, and all shells contribute no less than $-2A_2\overline{\psi}\beta_{\text{sh}}$. Their sum is the quantity inside (105), proving (106). Condition (107) then leaves a positive running margin after the field-drift and exact-remainder penalties of Section C.5. □

C.8.3 Finite-width attainability

To state the non-vacuity result independently of finite-width approximation errors, fix a constant $\psi_\star > 0$ such that the realized pretrained model satisfies $\psi_{\min}(0) \ge \psi_\star$, and define the ideal infinite-width, zero-movement support–leakage slack

$$d_\infty := \min_{\epsilon \in \{-1,+1\}} \left\{ \psi_\star p_\epsilon A_2 k_{\text{in}} - \overline{\psi} A_2 \left[2\beta_{\text{sh}} + k_{\text{out}} m_{-\epsilon}\right] \right\}. \tag{108}$$

Theorem C.12 (Attainability of a corrective finite-width regime)**.** *Assume $d_\infty > 0$ and fix $\delta_S \in (0,1)$. For fixed problem constants, there exist $\bar{\tau}_S > 0$, $\bar{\Lambda} > 0$, and a finite width m_0 such that, for every $m_h \geq m_0$, with probability at least $1 - \delta_S$ over the random hidden initialization, the following implication holds simultaneously for every source-training path satisfying the movement bound: if $\tau_S \leq \bar{\tau}_S$, the realized source model satisfies $\psi_{\min}(0) \geq \psi_\star$, and the target movement budget satisfies $\Lambda \leq \bar{\Lambda}$, then the certified pre-adaptation running margin is strictly positive. Equivalently, with d_{prim} used as the lower bound on d_0,*

$$d_{\mathrm{prim}} - 6\overline{\psi}\Gamma_0(\Lambda) - \frac{A_2}{2}\sigma_{\mathrm{drift}} > 0. \tag{109}$$

Consequently, every full-batch entropy-minimization schedule satisfying $\eta(T+1) \leq \Lambda$ obeys the early risk-descent conclusion of Theorem C.6.

Proof. On the common VC event, $v_m(\delta_S) \to 0$ as $m_h \to \infty$, and $b_m(t) \leq \sqrt{2/\pi}\, t + v_m(\delta_S)$. Hence Lemma C.9 implies that $\varepsilon_{\mathrm{gate}}$ can be made arbitrarily small by choosing τ_S sufficiently small and then taking m_h sufficiently large. Lemma C.10 shows that the same choices, together with a sufficiently small Λ, make $\Gamma_0(\Lambda)$ arbitrarily small. Finally, $\sigma_{\mathrm{drift}} = 2\Lambda A_2 \overline{\psi} \to 0$ as $\Lambda \to 0$.

The quantity d_{prim} differs from d_∞ only through the finite-width gate-overlap error $\varepsilon_{\mathrm{gate}}$ and the replacement of $\psi_\star$ by the no-smaller realized value $\psi_{\min}(0)$. Since $d_\infty > 0$, choose $\bar{\tau}_S$ and $\bar{\Lambda}$ small enough, and then m_0 large enough, that on the VC event $d_{\mathrm{prim}} \geq 3d_\infty/4$ and

$$6\overline{\psi}\Gamma_0(\Lambda) + \frac{A_2}{2}\sigma_{\mathrm{drift}} \leq \frac{d_\infty}{2}.$$

The left-hand side of (109) is then at least $d_\infty/4 > 0$. The VC event has probability at least $1 - \delta_S$, and the source-movement argument is uniform over all source-training paths satisfying the stated budget. □

The theorem separates two logically distinct requirements. Random initialization and bounded source movement control the finite-width coupling and the density of gates near ReLU boundaries. Correct semantic orientation of the source cores remains a property of the realized pretrained model, expressed through $\psi_{\min}(0) \geq \psi_\star$. No condition in the theorem refers to future target risk or to the success of the entropy-minimization trajectory.

D Additional Finite-Width Entropy Experiments

We examine whether source-informed component structure can produce correcting entropy dynamics in a naturally trained finite-width network. The study audits the proposed mechanism through realized trajectories, same-partition comparisons, and direct interventions on the initial field. Component assignments are constructed from source information and frozen before target labels are revealed. Target labels subsequently verify purity, semantic orientations, and risk; entropy updates receive neither labels nor orientations.

D.1 Natural high-dimensional setup

CIFAR-10 images are represented by the 640-dimensional penultimate embedding of a frozen RobustBench Standard WRN-28-10. For each binary source pair, we subtract the clean-source pair mean and apply ℓ_2 normalization. A two-layer ReLU network has 512 hidden units with bias, standard `torch.nn.Linear.reset_parameters` initialization, and an output head with independent entries $\pm 1/\sqrt{512}$ that remains fixed throughout training and adaptation. Source training minimizes binary logistic loss using SGD with learning rate 0.3, weight decay 0.03, and batch size 1024. We use class pairs $0/7$ and $1/6$ and source checkpoints at epochs 1 and 20.

Source-only K-means defines 16 components per class, and target samples are assigned to their nearest source prototypes before label inspection. Full-batch entropy minimization acts on complete two-component unions. Across the two pairs, two checkpoints, 13 corruptions, and 42 corruption–severity cells, the audit retains 1,018 semantically pure unions with at least one initial error. This construction tests correction in the original high-dimensional embedding without scalar logit projection, PCA, artificial features, or prescribed neuron roles. Purity is a criterion for inclusion in this audit pool, rather than an estimated prevalence over arbitrary target components.

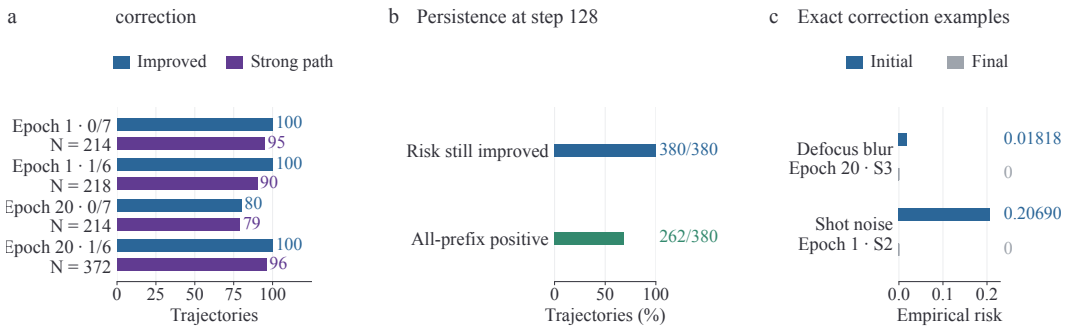

Figure 7: Finite-width entropy correction. (a) Recorded strict improvements and strong paths across the audited trajectory grid. (b) Persistence of the 380 improving trajectories through step 128. (c) Two representative exact trajectories showing complete correction of initial target errors. All updates are computed with the exact finite-width network and moving ReLU gates.

D.2 Corrective Trajectories and Exact Finite-Width Verification

To demonstrate that correcting trajectories exist across diverse optimization regimes, we evaluate candidate trajectories under a pre-defined grid of learning rates $\eta \in \{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 10^{-1}\}$ and audited horizons $T \in \{1, 2, 4, 8, 16, 32, 64, 128\}$. Across these configurations, we identify 380 trajectories exhibiting strict risk improvements, among which 360 strictly adhere to the strong-path condition—maintaining a positive realized field and positive signed score updates at every intermediate step. These improvements consistently emerge across both evaluated class pairs and source checkpoints: epoch-1 checkpoints yield 100 improving trajectories per pair, while epoch-20 checkpoints yield 80 and 100 for pairs $0/7$ and $1/6$, respectively (Figure 7a). This confirms that the error-correcting mechanism reliably manifests across distinct initialization states.

Across the 360 strong paths, the minimum realized field reaches 1.3524×10^{-5}, and the minimum signed single-step update is 3.8743×10^{-7}. The maximum numerical discrepancy in the exact score-update identity remains zero within machine precision, verifying that the finite-width network faithfully implements the theoretical positive-update mechanism. Furthermore, continuing all 380 improving trajectories up to step 128 reveals that risk remains strictly below baseline in all $380/380$ cases, even though only $262/380$ maintain positive signed updates across every evaluated prefix (Figure 7b). Hence, aggregate risk reduction remains robust within this horizon, persisting even when the stricter pointwise update criterion is locally violated.

D.3 Static–Dynamic Alignment on the Same Partition

Evaluated on the same component partition, the exact component Jacobian margin achieves a Spearman correlation of 0.6721 with the exact initial field d_0 (block-bootstrapped 95% confidence interval $[0.5864, 0.7776]$). In contrast, the balanced generic static margin correlates substantially more weakly, yielding 0.4053 with the Jacobian margin and 0.4547 with d_0. This pronounced alignment of the optimizer-aware margin corroborates the theoretical link between component-level support minus leakage and pointwise correcting forces, while demonstrating that a simple source-affinity margin cannot capture the dynamic coupling induced by entropy gradients.

The audited component-to-pointwise lower bound is $\underline{d}_0 := \Delta_J - \omega_J$, including the within-component oscillation correction in Proposition C.3. Double-precision recomputation on the same frozen unions with fixed output heads verifies $\Delta_J - \omega_J \le d_0$ in $1{,}018/1{,}018$ cases at absolute tolerance 2×10^{-14}; the maximum positive residual $[(\Delta_J - \omega_J) - d_0]_+$ is 1.91×10^{-17}. This verifies the numerical implementation of the deterministic lower bound. On the same audit pool, component support $\Delta_J > 0$ holds in 685 unions, the corrected lower bound is positive in 262, and the exact field d_0 is positive in 392. At $\Lambda = 10^{-3}$, 133 unions additionally satisfy the complete condition $\Delta_{\mathrm{corr}}(\Lambda) > 0$. Since the stability penalties and ω_J are nonnegative, these conditions obey $\Delta_{\mathrm{corr}}(\Lambda) \le \Delta_J - \omega_J \le d_0$ and $\Delta_J - \omega_J \le \Delta_J$.

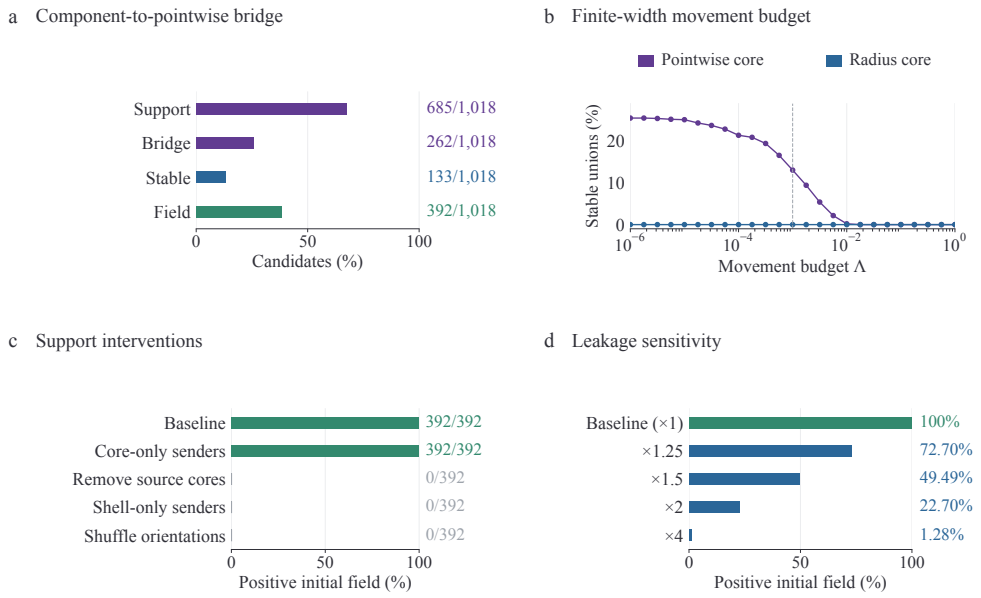

Figure 8: Pre-adaptation finite-width audit and direct field interventions. (a) Among 1,018 frozen unions, 685 satisfy component support $\Delta_J > 0$, 262 satisfy the component-to-pointwise bridge $\Delta_J - \omega_J > 0$, 133 satisfy the complete running-margin condition $\Delta_{\mathrm{corr}}(\Lambda) > 0$ at $\Lambda = 10^{-3}$, and 392 satisfy the exact initial-field condition $d_0 > 0$. (b) Sensitivity of the complete running-margin pass count to the movement budget Λ. (c) Source-support and orientation interventions among the 392 unions with baseline $d_0 > 0$. (d) Fraction retaining $d_0 > 0$ when cross-semantic coupling is multiplied by 1, 1.25, 1.5, 2, and 4. Panels (c,d) evaluate the initial field algebraically under receiver-fixed interventions; they are not adaptation-accuracy experiments.

The initial field also closely tracks the archived corrective trajectories. Across the recorded trajectory set, d_0 has Spearman correlation 0.7918 with realized risk reduction and achieves AUROC 0.9985 for distinguishing strictly improving trajectories, with $(\mathrm{TP}, \mathrm{FP}, \mathrm{FN}, \mathrm{TN}) = (380, 12, 0, 626)$. Together with the exact-update verification, this supports the role of the initial semantic field as a mechanistic indicator of whether the finite-width network admits a corrective entropy trajectory.

D.4 Direct interventions on support and leakage

To systematically isolate the constituent terms in the field decomposition, we hold the receiver set fixed and selectively perturb sender contributions and cross-semantic coupling across the 392 instances satisfying $d_0 > 0$. Retaining solely source-oriented core senders preserves a positive field in all $392/392$ cases. Conversely, ablating these cores, retaining only shell senders, or randomly permuting semantic orientations reduces the positive-field count to $0/392$ in every scenario (Figure 8c). These counterfactual evaluations verify that source-oriented cores constitute the decisive positive support governing these instances.

Selectively amplifying cross-semantic coupling induces a clear monotonic degradation: the fraction of instances maintaining a positive field drops from 100% at baseline to 72.70%, 49.49%, 22.70%, and 1.28% under scaling multipliers of 1.25, 1.5, 2, and 4, respectively (Figure 8d). This systematic decline directly corroborates the theoretical tension between positive support and semantic leakage. Importantly, these interventions evaluate the exact initial field decomposition algebraically, targeting field sign preservation rather than downstream empirical accuracy from re-optimized networks.

At the common pre-specified movement budget $\Lambda = 10^{-3}$, $133/1{,}018$ audited unions satisfy the complete running-margin condition $\Delta_{\mathrm{corr}}(\Lambda) > 0$. This is the full pre-adaptation certificate used by Theorem 4.1, rather than the weaker exact-field condition $d_0 > 0$, which holds on $392/1{,}018$ unions. Among these 133 certificate-positive unions, 11 additionally satisfy the theorem's shallow-error condition under an admissible full-batch schedule, so Theorem 4.1 certifies strict empirical-risk

reduction on $11/133$ of the fully certified cases ($11/1{,}018$ of the retained pool), rather than only positive signed-margin movement.

D.5 The pre-adaptation certificate and the realized field

The main theorem uses the pre-adaptation condition $\Delta_{\mathrm{corr}}(\Lambda) > 0$, which combines source-oriented support, within-component variation, and finite-width stability penalties. The exact initial field d_0 is the pointwise quantity entering the score dynamics; the component-to-pointwise bound and movement controls connect the pre-adaptation certificate to corrective score updates over the allowed horizon. The empirical audit therefore reports both the 133 unions satisfying the complete certificate and the 392 with $d_0 > 0$ as distinct properties of the same frozen pool.

Appendix C.8 gives an initialization-and-geometry construction as one explicit sufficient route to this regime. It derives positive support and controlled finite-width penalties from primitive source-oriented structure, without assuming future adaptation success. The main certificate itself is evaluated on the realized source-trained network and does not require that network to retain an isotropic gate law.

D.6 Representative exact trajectories

To inspect the underlying mechanics at sample resolution, we examine individual trajectories under finite-width dynamics. For the epoch-20 model evaluated on pair $0/7$ under severity-3 defocus blur, the designated 55-sample union contains a single initial error. Entropy minimization under a fixed learning rate $\eta = 0.1$ fully corrects this error within 8 steps, eliminating empirical risk from 0.01818 to zero (Figure 7c). Throughout this progression, both the initial and minimum realized fields equal 0.0068108, with a minimum signed score update of 0.0006812. Although the minimum oriented exact remainder is negative (-0.0005006), the numerical error of the exact-update identity remains zero within machine precision. This trajectory demonstrates that error correction proceeds successfully even in the presence of a negative finite-width remainder, highlighting the necessity of accounting for realized updates and higher-order remainders jointly.

The most pronounced risk reduction occurs for pair $0/7$ at source epoch 1 under severity-2 shot noise. In a 29-sample union, all six initial misclassifications are fully rectified within 64 optimization steps, driving risk from 0.20690 to zero while sustaining a minimum realized field of 0.0055781 and zero identity error. These trajectories provide concrete instances of complete error elimination governed by finite-width dynamics. We emphasize that these paths serve as constructive demonstrations of existence rather than estimators of average deployment gain.

D.7 Coverage of identified positive trajectories

The 380 identified strict-improvement trajectories collectively span 13 distinct corruption types across 42 corruption–severity conditions (Figure 9c). Among these, defocus blur yields 104 trajectories, shot noise 68, and Gaussian noise 39, with remaining cases distributed across motion and glass blur, environmental corruptions, brightness shifts, pixelation, and elastic deformations. This breadth demonstrates that the correcting mechanism operates reliably under a wide spectrum of visual distribution shifts. Because candidate sample sizes and evaluation budgets vary across specific corruption cells, these counts illustrate the empirical coverage of the theoretical mechanism rather than marginal success rates per perturbation.

E Posterior Calibration Bridge from PAC Identification to PRISM

This section strengthens the connection between the recoverable-semantic-structure theory and the theory-guided design of PRISM at the level of its actual probabilistic output. The key observation is that the same identification margin extends from hard component assignments to probabilistic posteriors. This yields a predictor-level calibration theorem and allows the posterior emitted by PRISM to be evaluated directly against the same semantic identification objective on a frozen component structure.

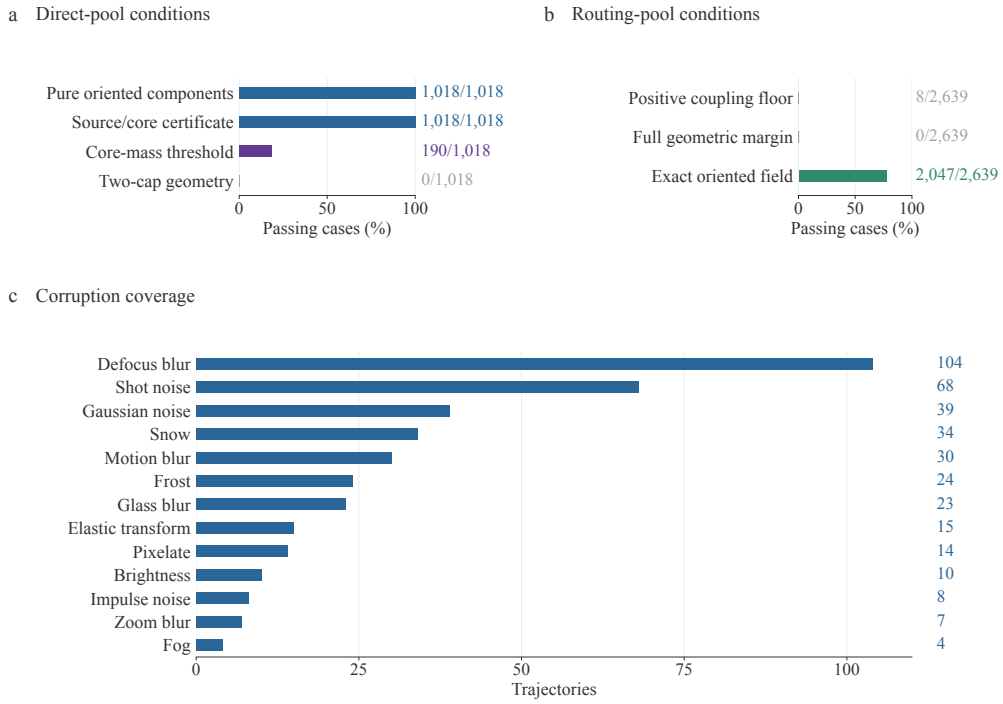

Figure 9: Sufficient-condition verification and corruption coverage. (a) Verification rates across the 1,018 evaluated candidates, where purity serves as an inclusion criterion. (b) Gate-overlap and routing properties across a baseline pool of 2,639 candidate configurations. The two panels evaluate distinct populations. (c) Distribution of the 380 identified strict-improvement trajectories across 13 corruption types, ordered by frequency. These values reflect coverage across 42 corruption–severity regimes rather than per-corruption success probabilities.

E.1 From hard component labels to probabilistic outputs

Recall the component graph $\{C_u\}_{u\in\mathcal{V}}$, masses m_u, affinities w_{uv}, source-anchor masses a_u^j, and semantic component labels $c(u)$ from Section 3. Let $\Delta_K := \{q \in \mathbb{R}_+^K : \sum_j q_j = 1\}$, and assign each component a posterior $\pi_u \in \Delta_K$. Writing $\Pi = (\pi_u)_{u\in\mathcal{V}}$, define

$$\mathcal{J}_{\mathrm{TV}}(\Pi) := \frac{1}{2}\sum_{u<v} w_{uv}\|\pi_u - \pi_v\|_1 + \lambda \sum_{u\in\mathcal{V}}\sum_{j=1}^{K} a_u^j(1-\pi_{u,j}). \tag{110}$$

For every discrete assignment y, taking $\pi_u = e_{y_u}$ gives $\mathcal{J}_{\mathrm{TV}}((e_{y_u})_u) = \mathcal{J}(y)$: the total-variation distance between two distinct one-hot vectors is one, and the anchor term becomes the original wrong-anchor penalty.

Let $\Pi^\star := (e_{c(u)})_u$ be the semantic one-hot assignment and define $M_{\mathrm{soft}}(\Pi) := \sum_{u\in\mathcal{V}} m_u(1 - \pi_{u,c(u)})$.

Theorem E.1 (Posterior calibration)**.** *Under Assumption 3.1, every* $\Pi \in \Delta_K^{|\mathcal{V}|}$ *satisfies*

$$\mathcal{J}_{\mathrm{TV}}(\Pi) - \mathcal{J}_{\mathrm{TV}}(\Pi^\star) \geq \Delta_{\mathrm{id}} M_{\mathrm{soft}}(\Pi). \tag{111}$$

Consequently, if $\Delta_{\mathrm{id}} > 0$*, then* $\Pi^\star$ *is the unique global minimizer of* $\mathcal{J}_{\mathrm{TV}}$.

Proof. Write $q_u := 1 - \pi_{u,c(u)} \in [0,1]$. If $c(u) = c(v) = k$, total variation dominates every coordinate difference, so $\frac{1}{2}\|\pi_u - \pi_v\|_1 \geq |\pi_{u,k} - \pi_{v,k}| = |q_u - q_v|$. If $c(u) \neq c(v)$, the semantic one-hot endpoints have total variation one; thus the triangle inequality gives $\frac{1}{2}\|\pi_u - \pi_v\|_1 - 1 \geq -(q_u + q_v)$.

For $k = c(u)$, the anchor increment relative to $\Pi^\star$ is

$$\lambda \left[a_u^k q_u - \sum_{j \neq k} a_u^j \pi_{u,j} \right] \geq \lambda \left[a_u^k - \max_{j \neq k} a_u^j \right] q_u, \tag{112}$$

using $\sum_{j \neq k} \pi_{u,j} = q_u$.

For $t \in [0,1]$, define the level set $S_{k,t} := \{u \in \mathcal{V}_k : q_u > t\}$. The connected-set A3 condition extends to an arbitrary $S_{k,t}$ by decomposing it into connected components. Hence the layer-cake/coarea identity gives

$$\sum_{\substack{u<v \\ c(u)=c(v)=k}} w_{uv} |q_u - q_v| + \lambda \sum_{u \in \mathcal{V}_k} a_u^k q_u \tag{113}$$

$$= \int_0^1 \left[\partial_k(S_{k,t}) + \lambda A_k^{\mathrm{corr}}(S_{k,t}) \right] dt \geq \kappa_{\mathrm{sup}} \sum_{u \in \mathcal{V}_k} m_u q_u. \tag{114}$$

Summing over k yields support at least $\kappa_{\mathrm{sup}} \sum_u m_u q_u$. All cross-semantic edges and wrong anchors can reduce the objective by at most

$$\sum_u q_u \left[\sum_{v : c(v) \neq c(u)} w_{uv} + \lambda \max_{j \neq c(u)} a_u^j \right] \leq \rho_{\mathrm{leak}} \sum_u m_u q_u. \tag{115}$$

Subtracting leakage from support proves (111). If $\Delta_{\mathrm{id}} > 0$, equality at the global minimum forces $M_{\mathrm{soft}}(\Pi) = 0$, and thus $\pi_u = e_{c(u)}$ for every u. □

E.2 A RISK BOUND FOR ANY PROBABILISTIC TTA PREDICTOR

Let $p : \mathcal{X} \to \Delta_K$ be any probabilistic predictor. It need not be constant inside a component and need not minimize $\mathcal{J}_{\mathrm{TV}}$. Define

$$\pi_u(p) := \frac{1}{m_u} \int_{C_u} p(x)\, d\mu(x), \qquad \Pi(p) := (\pi_u(p))_u, \tag{116}$$

and the *realization gap*

$$e(p) := \mathcal{J}_{\mathrm{TV}}(\Pi(p)) - \min_{\Pi \in \Delta_K^{|\mathcal{V}|}} \mathcal{J}_{\mathrm{TV}}(\Pi). \tag{117}$$

Corollary E.2 (Posterior-to-risk bridge)**.** *If $\Delta_{\mathrm{id}} > 0$ and $h_p(x) := \arg\max_j p_j(x)$, then*

$$L_Q(h_p) \leq \min\left\{ 1,\ \zeta + \alpha_{\mathrm{tot}} + \frac{2e(p)}{\Delta_{\mathrm{id}}} \right\}. \tag{118}$$

Proof. Theorem E.1 gives $M_{\mathrm{soft}}(\Pi(p)) \leq e(p)/\Delta_{\mathrm{id}}$. For $x \in C_u$, if $h_p(x) \neq c(u)$ then $p_{c(u)}(x) \leq 1/2$, and therefore $\mathbf{1}\{h_p(x) \neq c(u)\} \leq 2(1 - p_{c(u)}(x))$. Hence

$$L_Q(h_p) \leq \zeta + \sum_u Q(Y \neq c(u), X \in C_u) + \sum_u \int_{C_u} \mathbf{1}\{h_p(x) \neq c(u)\}\, d\mu(x) \tag{119}$$

$$\leq \zeta + \alpha_{\mathrm{tot}} + 2M_{\mathrm{soft}}(\Pi(p)) \leq \zeta + \alpha_{\mathrm{tot}} + \frac{2e(p)}{\Delta_{\mathrm{id}}}. \tag{120}$$

Clipping by one completes the proof. □

Because the calibration is predictor-agnostic, it applies to the source posterior, naive multi-view averaging, and PRISM under the same fixed component graph. It therefore provides a common semantic scale on which these practical realizations can be compared through their realization gaps.

E.3 Stage-2 computation as a linear program

The relaxed identification problem is convex and is a linear program. For every nonzero edge (u, v) and class j, introduce $t_{uvj} \geq |\pi_{u,j} - \pi_{v,j}|$. Then

$$\begin{aligned} \min_{\Pi, t} \quad & \frac{1}{2} \sum_{u<v} \sum_{j} w_{uv} t_{uvj} + \lambda \sum_{u,j} a_u^j (1 - \pi_{u,j}) \\ \text{s.t.} \quad & t_{uvj} \geq \pi_{u,j} - \pi_{v,j}, \qquad t_{uvj} \geq \pi_{v,j} - \pi_{u,j}, \\ & \pi_{u,j} \geq 0, \qquad \sum_{j} \pi_{u,j} = 1. \end{aligned} \tag{121}$$

Thus Stage 2 posterior identification is polynomial-time solvable. Under $\Delta_{\mathrm{id}} > 0$, Theorem E.1 further shows that its unique population optimum is the semantic one-hot assignment. This statement concerns Stage 2 only and does not remove the component-estimation requirement in A2.

Proposition E.3 (Empirical integrality under a positive identification margin)**.** *Fix an empirical component system and let $\widehat{m}_u$, $\widehat{w}_{uv}$, and $\widehat{a}_u^j$ denote its empirical component masses, edge weights, and anchor masses. Let $\widehat{\mathcal{J}}_{\mathrm{TV}}$ be the linear-program objective obtained from (121) by replacing the population coefficients with these empirical quantities. Using the semantic component assignment c only for analysis, define the empirical support and leakage constants $\widehat{\kappa}_{\mathrm{sup}}$ and $\widehat{\rho}_{\mathrm{leak}}$ exactly as in A3–A4, and let*

$$\widehat{\Delta}_{\mathrm{id}} := \widehat{\kappa}_{\mathrm{sup}} - \widehat{\rho}_{\mathrm{leak}}.$$

If $\widehat{\Delta}_{\mathrm{id}} > 0$, then every $\Pi \in \Delta_K^{|\widehat{\mathcal{V}}|}$ satisfies

$$\widehat{\mathcal{J}}_{\mathrm{TV}}(\Pi) - \widehat{\mathcal{J}}_{\mathrm{TV}}(\Pi^\star) \geq \widehat{\Delta}_{\mathrm{id}} \sum_{u} \widehat{m}_u \big(1 - \pi_{u,c(u)}\big),$$

where $\Pi^\star = (e_{c(u)})_u$. Consequently, $\Pi^\star$ is the unique optimum of the empirical LP and is integral. No rounding step is required.

Proof. The proof is the empirical analogue of Theorem E.1. Write $q_u := 1 - \pi_{u,c(u)}$. For each class k and threshold t, apply the empirical A3 inequality to the level set $S_{k,t} := \{u : c(u) = k,\ q_u > t\}$. The same layer-cake identity lower-bounds the empirical within-semantic and correct-anchor contribution by $\widehat{\kappa}_{\mathrm{sup}}$ times the empirical soft mismatch mass. Empirical cross-semantic edges and wrong anchors can reduce the objective by at most $\widehat{\rho}_{\mathrm{leak}}$ times the same mass. Subtracting the two terms yields the displayed inequality. Strict positivity of $\widehat{\Delta}_{\mathrm{id}}$ forces zero soft mismatch at every optimum. □

The LP itself uses only the empirical graph and anchor coefficients; the semantic labels $c(u)$ enter only the analysis of when its optimum is certifiably integral. Thus, whenever the empirical positive-margin event holds, Stage 2 semantic identification is polynomial-time solvable and the LP returns exactly the same semantic assignment certified by the discrete margin analysis.

E.4 Module-specific guarantees for semantic restoration

We give sufficient conditions under which the particular operations of PRISM restore source-oriented evidence and decrease the fixed-graph realization gap. These conditions concern the geometry of target shift and the semantic orientation of repeated measurements; they are additional to, and are not implied by, Assumption 3.1. Throughout, the component partition, graph weights, and source anchors in $\mathcal{J}_{\mathrm{TV}}$ remain fixed. In particular, an operation cannot obtain a smaller gap merely by changing the graph against which it is evaluated. We assume $K \geq 2$, $r \geq 1$, and finite second moments. Population statements have the corresponding deterministic interpretation on a fixed empirical target pool. We use column vectors, so the corrected logits are $AG_n(z_v - \widehat{\mu}_v) + b$, equivalently the transpose convention in Section 5.

Classifier visibility is an exact reduction, not a source of accuracy.

Proposition E.4 (Classifier-visible sufficiency)**.** *Let $P_K = I - K^{-1}\mathbf{1}\mathbf{1}^\top$, $W_c = P_K W$, and let the columns of B form an orthonormal basis of the row space of W_c. With $A = W_c B$, every feature h satisfies* $\mathrm{softmax}(Wh + b) = \mathrm{softmax}(AB^\top h + b)$. *Consequently, every feature correction has a posterior-equivalent correction in* $\mathrm{span}(B)$*, and arbitrary variation in* $\ker(W_c)$ *cannot contaminate the visible-space moments used by* PRISM.

Proof. Since $W_c B B^\top = W_c$, the logits Wh and $AB^\top h$ differ only by a class-shared scalar. Softmax removes this scalar. Applying the same identity to a corrected feature proves the second statement. □

A source-oriented model of target distortion. Consider the population decomposition

$$z_v = d_v + Ht + \epsilon_v, \qquad \mathbb{E}t = 0, \quad \mathbb{E}\epsilon_v = 0, \quad \mathbb{E}[t\epsilon_v^\top] = 0, \tag{122}$$

where $t \in \mathbb{R}^r$ is the latent task representation, d_v is a systematic view offset, and $H \succ 0$ is a shared, source-coordinate-preserving anisotropic distortion. Write $\Sigma_t = \mathbb{E}[tt^\top]$ and $R_v = \mathbb{E}[\epsilon_v \epsilon_v^\top]$. Different views may have correlated noise. Semantic orientation enters through the frozen head: for some $\gamma > 0$, assume that, almost everywhere on every resolved component,

$$(A_{c(u),:} - A_{k,:})t(x) + b_{c(u)} - b_k \geq \gamma, \qquad x \in C_u, \quad k \neq c(u). \tag{123}$$

Thus the source head correctly orients the latent representation, not necessarily its distorted observations. Neither t, H, nor the labels $c(u)$ are supplied to adaptation. The zero-mean condition fixes the latent origin relative to the unchanged classifier bias; it does not assert that centering is harmless under arbitrary class-prior changes.

For a fixed linear transform M, define the logit-restoration error $\mathcal{E}_v(M, d) := \mathbb{E}\|A[M(z_v - d) - t]\|_2^2$. All moment identities below concern fixed population transforms. We handle the data-dependent, current-batch-inclusive transforms separately.

Proposition E.5 (View-wise centering removes offset error)**.** *Under* (122)*,*

$$\mathcal{E}_v(M, d_v) = \mathrm{Tr}\big(A\big[(MH - I)\Sigma_t(MH - I)^\top + MR_vM^\top\big]A^\top\big), \tag{124}$$

$$\mathcal{E}_v(M, 0) - \mathcal{E}_v(M, d_v) = \|AMd_v\|_2^2. \tag{125}$$

Hence exact view-wise centering removes precisely the classifier-visible squared offset, without changing the residual covariance error.

Proof. The centered residual equals $A[(MH - I)t + M\epsilon_v]$ and has mean zero. Its two terms are uncorrelated by (122). Without centering, the same residual acquires the deterministic mean AMd_v. Expanding its squared norm gives both identities. □

If instead $\mathbb{E}t \neq 0$, population centering subtracts $d_v + H\mathbb{E}t$ and introduces the additional semantic displacement $-AMH\mathbb{E}t$ relative to the original latent coordinates. Thus centering has a semantic justification when the estimated mean represents a nuisance offset, not merely because that mean is observable.

Total-covariance transport reverses source-aligned anisotropy.

Proposition E.6 (Restoration by the actual shrinkage and damping rule)**.** *Suppose $\epsilon_v = 0$, $\Sigma_t = \sigma^2 I$ with $\sigma > 0$, and $H = U\,\mathrm{diag}(h_1, \ldots, h_r)U^\top \succ 0$ with $r^{-1}\,\mathrm{Tr}(H^2) = 1$. Construct T from the population total covariance using the same shrinkage and one-sided damping as* PRISM*, allowing any $\rho \in [0, 1]$ and $\kappa \in [0, 1]$. Its eigenvalues are*

$$a_j = \begin{cases} [(1-\rho)h_j^2 + \rho]^{-\kappa/2}, & h_j < 1, \\ [(1-\rho)h_j^2 + \rho]^{-1/2}, & h_j \geq 1. \end{cases} \tag{126}$$

Then $h_j \leq a_j h_j \leq 1$ when $h_j \leq 1$, and $1 \leq a_j h_j \leq h_j$ when $h_j \geq 1$. Consequently,

$$\mathcal{E}_v(T, d_v) = \sigma^2 \sum_j \|AU_{:j}\|_2^2 (a_j h_j - 1)^2 \leq \mathcal{E}_v(I, d_v). \tag{127}$$

With $\rho = 0$ and $\kappa = 1$, $T = H^{-1}$ and restoration is exact.

Proof. Here $S_T = \sigma^2 H^2$ and its average eigenvalue is σ^2. Shrinkage replaces h_j^2 by $(1-\rho)h_j^2 + \rho$, giving (126). For $h_j \geq 1$ this replacement lies in $[1, h_j^2]$, whereas for $h_j \leq 1$ it lies in $[h_j^2, 1]$. The damping exponent preserves these intervals. Thus each restored multiplier lies between its uncorrected value and one. Substitute into (124). □

Example E.7 (Strict correction of an anisotropy-induced error)**.** Let $K = 3$, $r = 2$, $b = 0$, and let the latent representation be uniform on $t_1 = (1, 0)^\top$, $t_2 = (-1/2, \sqrt{3}/2)^\top$, and $t_3 = (-1/2, -\sqrt{3}/2)^\top$, with labels $1, 2, 3$. Take $A_{k,:} = t_k^\top$, no offset or noise, and $H = 17^{-1/2} \left(\begin{smallmatrix} 1 & 2 \\ 2 & 5 \end{smallmatrix}\right)$. Then $\mathbb{E}t = 0$, $\Sigma_t = I/2$, $\mathrm{Tr}(H^2)/2 = 1$, and the latent head margin is $3/2$. Nevertheless, Ht_1 is predicted as class 2, because $t_2^\top Ht_1 = (\sqrt{3} - 1/2)/\sqrt{17} > 1/\sqrt{17} = t_1^\top Ht_1$. The full population transport $T = H^{-1}$ restores all three labels exactly. Thus covariance correction can remove a genuine decision error, not just reduce a moment discrepancy. The corrected strict margins persist under sufficiently small damping, shrinkage, and estimation errors, where damping error means deviation of κ from one.

The trace normalization separates anisotropic shape from an unidentified common scale. Likewise, isotropy of the latent covariance is a substantive compatibility condition, not something established by whitening the target. An arbitrary orthogonal semantic rotation has no covariance signature that this operation can invert. The proposition uses the implemented form of shrinkage, but does not require an optimality claim for OAS (Chen et al., 2010) on correlated repeated measurements.

The noisy case makes the amplification tradeoff explicit. Suppose H, Σ_t, and R_v share an orthonormal eigenbasis, with eigenvalues h_j, s_j, and ν_{vj}, respectively. A multiplier $g > 0$ in direction j contributes $\|AU_{:j}\|_2^2 F_{vj}(g)$, where

$$\begin{aligned} F_{vj}(g) &:= (gh_j - 1)^2 s_j + g^2 \nu_{vj}, \\ F_{vj}(g) - F_{vj}(1) &= (g-1)\big[(g+1)(h_j^2 s_j + \nu_{vj}) - 2h_j s_j\big]. \end{aligned} \tag{128}$$

If $h_j \geq 1$ and $g \in [1/h_j, 1]$, both the distortion and noise contributions are nonincreasing. For an expansion $g > 1$, strict improvement holds exactly when $\nu_{vj} < s_j[2h_j/(g+1) - h_j^2]$. This is an explicit signal-to-nuisance condition: small total variance alone does not justify amplification. The finite-prefix exponent reduces the expansion and its associated noise penalty, but cannot create a missing semantic signal.

Within-orbit suppression has a quantifiable signal cost. Let $\bar{\hat{z}}_i = V^{-1} \sum_v (z_{iv} - \hat{\mu}_v)$. The sample quantities used by PRISM satisfy the exact identity

$$S_W = S_T - S_B = \frac{1}{nV} \sum_{i,v} (\hat{z}_{iv} - \bar{\hat{z}}_i)(\hat{z}_{iv} - \bar{\hat{z}}_i)^\top \succeq 0. \tag{129}$$

Under (122), its population counterpart is

$$S_W^{\mathrm{pop}} = \frac{1}{V} \sum_v \mathbb{E}[(\epsilon_v - \bar{\epsilon})(\epsilon_v - \bar{\epsilon})^\top] = \frac{1}{V} \sum_v R_v - \mathrm{Cov}(\bar{\epsilon}), \qquad \bar{\epsilon} = \frac{1}{V} \sum_v \epsilon_v. \tag{130}$$

Both identities follow by expanding around the view average; the shared Ht cancels before any covariance is taken. Independent noises with common covariance R give $S_W^{\mathrm{pop}} = (1 - 1/V)R$. Shared nuisance also cancels, so S_W identifies view-varying nuisance, not every semantically irrelevant factor.

Proposition E.8 (When nuisance suppression improves restoration)**.** *Suppose T, N, H, Σ_t, and R_v share an orthonormal eigenbasis. Write a_j for the multiplier of T and $c_j = \min\{\sqrt{\eta_{\mathrm{avg}}/\eta_j}, 1\}$ for that of N, with multiplier one on zero-variance directions. Then $G = N^{1/2} T N^{1/2} = NT$ has multiplier $a_j c_j$, and*

$$F_{vj}(a_j c_j) - F_{vj}(a_j) = (1 - c_j)\big[2a_j h_j s_j - (1 + c_j) a_j^2 (h_j^2 s_j + \nu_{vj})\big]. \tag{131}$$

For a contracted direction, suppression is beneficial precisely when $(1 + c_j) a_j (h_j^2 s_j + \nu_{vj}) > 2h_j s_j$. In particular, after exact anisotropy correction $a_j h_j = 1$, this becomes

$$\frac{\nu_{vj}}{h_j^2} > \frac{1 - c_j}{1 + c_j} s_j. \tag{132}$$

If the non-strict inequalities hold for every contracted direction, then $\mathcal{E}_v(G, d_v) \leq \mathcal{E}_v(T, d_v)$.

Proof. Subtract the two quadratic errors in (128). The resulting factorization is (131). Its sign yields the stated criterion; summing with nonnegative coefficients $\|AU_{:j}\|_2^2$ proves the final conclusion. □

Equation (132) compares the removed noise with the attenuated semantic signal. A direction containing only nuisance ($s_j = 0$) is safely contracted, whereas large within-orbit variance alone does not justify suppressing a strongly discriminative direction. The symmetric sandwich is not silently replaced by NT when the matrices do not commute. Indeed, with $[T, N^{1/2}] = TN^{1/2} - N^{1/2}T$,

$$\sqrt{\mathcal{E}_v(G, d_v)} \le \sqrt{\mathcal{E}_v(NT, d_v)} + \|A\|_{\mathrm{op}}\|[T, N^{1/2}]\|_{\mathrm{op}}\sqrt{\mathbb{E}\|z_v - d_v\|_2^2}. \tag{133}$$

To see this, use $G - NT = N^{1/2}[T, N^{1/2}]$, $\|N^{1/2}\|_{\mathrm{op}} \le 1$, and the triangle inequality in L_2. Thus noncommuting geometry is controlled by an explicit penalty rather than by positive definiteness of G alone.

From restoration to semantic risk and the actual realization gap. Define the fixed-graph sensitivity constants

$$\beta_u := \sum_{u' \ne u} w_{uu'} + \lambda\big(\max_j a_u^j - \min_j a_u^j\big), \qquad B_J := \max_u \frac{\beta_u}{m_u}, \qquad K_J := \sum_u \beta_u. \tag{134}$$

For any two component-posterior systems,

$$|\mathcal{J}_{\mathrm{TV}}(\Pi) - \mathcal{J}_{\mathrm{TV}}(\Pi')| \le \sum_u \beta_u \,\mathrm{TV}(\pi_u, \pi'_u). \tag{135}$$

The edge part follows from the triangle inequality for total variation. For the anchor part, the entries of $\pi_u - \pi'_u$ sum to zero, so its inner product with $(a_u^j)_j$ is bounded by the range of the anchor weights times total variation.

Theorem E.9 (Semantic restoration and strict gap improvement)**.** *Assume* (123). *Let $\ell_v(x)$ be the actual view logits, $r_v(x) := \ell_v(x) - (At(x) + b)$, and $p(x) = \sum_v w_v(x)\,\mathrm{softmax}(\ell_v(x))$, where the nonnegative weights sum to one and may depend arbitrarily on the observed posteriors. Put*

$$E_w := \mathbb{E}\sum_v w_v(X)\|r_v(X)\|_2^2, \qquad \nu_\gamma := \frac{(K-1)e^{-\gamma}}{1 + (K-1)e^{-\gamma}}, \qquad \chi_\gamma := \frac{1 - e^{-\gamma}}{1 + (K-1)e^{-\gamma}}, \tag{136}$$

and let $m_{\mathrm{res}} = 1 - \zeta > 0$. Then

$$M_{\mathrm{soft}}(\Pi(p)) \le m_{\mathrm{res}}\nu_\gamma + \frac{1}{2}\sqrt{m_{\mathrm{res}}E_w}, \tag{137}$$

$$L_Q(\arg\max p) \le \min\left\{1,\ \zeta + \alpha_{\mathrm{tot}} + \frac{E_w}{\chi_\gamma^2}\right\}. \tag{138}$$

The risk bound does not require a positive graph identification margin. If additionally $\Delta_{\mathrm{id}} > 0$ under the posterior-calibration conditions, then

$$e(p) \le B_J\left(m_{\mathrm{res}}\nu_\gamma + \frac{1}{2}\sqrt{m_{\mathrm{res}}E_w}\right). \tag{139}$$

Consequently, for any fixed-graph comparator p_0 satisfying $e(p_0)/B_J > m_{\mathrm{res}}\nu_\gamma$, the explicit condition

$$E_w < \frac{4}{m_{\mathrm{res}}}\left(\frac{e(p_0)}{B_J} - m_{\mathrm{res}}\nu_\gamma\right)^2 \tag{140}$$

implies $e(p) < e(p_0)$.

Proof. Write $q^* = \mathrm{softmax}(At + b)$. The latent margin implies $1 - q^*_{c(u)} \le \nu_\gamma$ and $q^*_{c(u)} - q^*_k \ge \chi_\gamma$ on C_u. For any logits ℓ and increment r, $\mathrm{TV}(\mathrm{softmax}(\ell + r), \mathrm{softmax}(\ell)) \le \|r\|_2/2$: along the interpolating path, the ℓ_1 norm of the softmax derivative is $\mathbb{E}_q|r_j - \mathbb{E}_q r_j| \le \|r\|_2$. Therefore $\mathrm{TV}(p, q^*) \le \frac{1}{2}\sum_v w_v\|r_v\|_2$. Integrating the correct-coordinate deficit and using Cauchy–Schwarz gives (137).

If $\arg\max p \neq c(u)$, some $k \neq c(u)$ satisfies $p_k \geq p_{c(u)}$, which forces $\chi_\gamma \leq 2\,\mathrm{TV}(p, q^*) \leq \sum_v w_v \|r_v\|_2$. The square of the last expression is at most $\sum_v w_v \|r_v\|_2^2$. Integration, with unresolved mass and minority labels charged separately, proves (138).

When $\Delta_{\mathrm{id}} > 0$, Theorem E.1 identifies $\Pi^* = (\mathbf{e}_{c(u)})_u$ as the minimizer. Applying (135) against Π^* gives $e(p) \leq \sum_u \beta_u (1 - \pi_{u,c(u)}(p)) \leq B_J M_{\mathrm{soft}}(\Pi(p))$. This proves (139); solving its strict comparison with $e(p_0)$ gives (140). □

Propositions E.5–E.8 identify primitive conditions reducing restoration error; the theorem states how much reduction suffices for an actual objective comparison. A smaller restoration error, or a smaller upper bound, alone does not imply $e(p) < e(p_0)$. For data-dependent fusion we likewise do not replace E_w by a fixed-weight average: the always-valid estimate is $E_w \leq \sum_v \mathbb{E}\|r_v\|_2^2$. Strict accuracy improvement follows when the right-hand side of (138) is below the comparator risk. In particular, when $\zeta = \alpha_{\mathrm{tot}} = 0$, if $\sum_v w_v(x)\|r_v(x)\|_2 < \chi_\gamma$ everywhere, the fused predictor is correct everywhere and strictly improves any comparator with positive risk.

Finite-prefix and arrival-time stability. Let G_0 be a fixed reference transform and let G_n, $\widehat{\mu}_v$ be the actual statistics available when a sample arrives. Pathwise,

$$\begin{aligned} &\|AG_n(z_v - \widehat{\mu}_v) - AG_0(z_v - d_v)\|_2 \\ &\quad \leq \|A\|_{\mathrm{op}} \left[\|G_n - G_0\|_{\mathrm{op}} \|z_v - d_v\|_2 + \|G_n\|_{\mathrm{op}} \|\widehat{\mu}_v - d_v\|_2\right]. \end{aligned} \tag{141}$$

This follows by adding and subtracting $AG_n(z_v - d_v)$. It transfers the population restoration calculations to actual emitted logits through an explicit estimation penalty, even when the sample being scored contributes to the current batch statistics. No claim that nV views are independent observations is used. Historical outputs remain fixed; changes of later transforms are not applied retrospectively.

Why squared spectral loading and local JS evidence can decrease the gap. To isolate fusion, consider an additional source-oriented posterior model. For this result take $\zeta = 0$, and assume at least two semantic classes have positive mass. Define reference view posteriors

$$q_v^0(x) = u_K + \theta_v(\mathbf{e}_{c(u)} - u_K), \qquad x \in C_u, \quad 0 \leq \theta_v \leq 1, \quad \sum_v \theta_v^2 > 0. \tag{142}$$

The loading θ_v measures correctly oriented evidence, not arbitrary shared variance. This semantic ray condition is not a consequence of the repeated-measurement decomposition. Perturbations of these reference posteriors are treated below. Write $c(x) = c(u)$ for $x \in C_u$, let $\varpi_k = \mu(\{x : c(x) = k\})$, and define $\sigma_{\mathrm{sem}}^2 = 1 - \|\varpi\|_2^2 > 0$ and $S_2 = \sum_v \theta_v^2$. Define $d(a) = \sqrt{\mathrm{JS}(u_K + a(\mathbf{e}_1 - u_K), u_K)}$. By permutation symmetry, the same function applies to every semantic class.

Proposition E.10 (Operation-specific descent of the fixed-graph gap)**.** *Under (142), the centered collision matrix is $C^0 = \sigma_{\mathrm{sem}}^2 \theta\theta^\top$, and the spectral leverage is exactly $L_v^* = \theta_v^2/S_2$. Equal fusion, spectral fusion, and full spectral–JS fusion have component posteriors $\pi_u^a = u_K + b_a(\mathbf{e}_{c(u)} - u_K)$, respectively, with*

$$\begin{aligned} b_{\mathrm{av}} = \frac{1}{V}\sum_v \theta_v, \quad b_{\mathrm{sp}} = \frac{\sum_v \theta_v^3}{S_2}, \quad b_{\mathrm{full}} = \frac{\sum_v \theta_v^3 d(\theta_v)}{\sum_v \theta_v^2 d(\theta_v)}, \\ b_{\mathrm{full}} \geq b_{\mathrm{sp}} \geq b_{\mathrm{av}}. \end{aligned} \tag{143}$$

Let $\Pi_U = (u_K)_u$ and define the aggregate graph–anchor alignment

$$\begin{aligned} \mathfrak{a}_{\mathrm{sem}} &:= \mathcal{J}_{\mathrm{TV}}(\Pi_U) - \mathcal{J}_{\mathrm{TV}}(\Pi^*) \\ &= \lambda \sum_u \left(a_u^{c(u)} - \frac{1}{K}\sum_j a_u^j \right) - \sum_{u<u':\, c(u)\neq c(u')} w_{uu'}. \end{aligned} \tag{144}$$

For any two of the three fusion stages,

$$e(p_b^0) - e(p_a^0) = -(b_b - b_a)\mathfrak{a}_{\mathrm{sem}}. \tag{145}$$

Hence $\mathfrak{a}_{\mathrm{sem}} > 0$ guarantees that both spectral leverage and local JS reweighting decrease the realization gap whenever their effective loading increases strictly. This comparison does not require $\Delta_{\mathrm{id}} > 0$. Under positive-margin posterior calibration, $\mathfrak{a}_{\mathrm{sem}} \geq \Delta_{\mathrm{id}}(1 - 1/K) > 0$ automatically.

Proof. Centering gives $q_v^0 - \mathbb{E}q_v^0 = \theta_v(\mathbf{e}_{c(x)} - \varpi)$. Its collision matrix is therefore $\sigma_{\text{sem}}^2\theta\theta^\top$, whose leading unit eigenvector is $\theta/\sqrt{S_2}$ up to sign. Substituting its squared coordinates into the two fusion rules gives the three loadings in (143). The inequality $b_{\text{sp}} \geq b_{\text{av}}$ follows from $\text{Cov}_{v\sim\text{Unif}}(\theta_v^2, \theta_v) \geq 0$.

The function d is increasing on $[0,1]$. Indeed, writing $r = \mathbf{e}_1 - u_K$, the second derivative of $J(a) = \text{JS}(u_K + ar, u_K)$ on $(0,1)$ is

$$J''(a) = \frac{1}{2}\sum_k \frac{(u_K)_k r_k^2}{((u_K)_k + ar_k)(2(u_K)_k + ar_k)} > 0, \qquad J'(0) = 0. \tag{146}$$

Thus $b_{\text{full}} - b_{\text{sp}} = \text{Cov}_{L^*}(\theta_v, d(\theta_v))/\mathbb{E}_{L^*}d(\theta_v) \geq 0$. For any increasing function f, the covariance used here is nonnegative because it equals $\frac{1}{2}\mathbb{E}[(\theta_v - \theta_{v'})(f(\theta_v) - f(\theta_{v'}))]$.

Along $\Pi(b) = (1-b)\Pi_U + b\Pi^*$, every cross-semantic edge has total variation b, every same-semantic edge has total variation zero, and the anchor term is affine. Hence $\mathcal{J}_{\text{TV}}(\Pi(b)) = \mathcal{J}_{\text{TV}}(\Pi_U) - b\mathfrak{a}_{\text{sem}}$. The same minimum is subtracted in both realization gaps, proving (145) even when Π^* is not the minimizer. Finally, apply posterior calibration to Π_U, whose soft mismatch is $1 - 1/K$, to obtain the positive-margin implication. □

For equal and purely spectral population fusion, the same conclusion extends to valid posteriors $q_v = q_v^0 + \xi_v \in \Delta_K$ with $\mathbb{E}[\xi_v \mid C_u] = 0$ and $\mathbb{E}[\xi_v^\top \xi_{v'}] = \sigma_\xi^2 \mathbf{1}\{v = v'\}$: then $C = \sigma_{\text{sem}}^2\theta\theta^\top + \sigma_\xi^2 I$ has the same leading direction, and the component averages are unchanged. Individual view predictions may therefore be incorrect.

This conditional-mean model itself follows, for example, from corrected logits $\ell_v = \gamma_v \mathbf{e}_{c(u)} + \omega_v \mathbf{1} + \varepsilon_v^\ell$, where $\gamma_v > 0$ and the noise law is independent of the component and invariant under permutations of class coordinates. At $\gamma_v = 0$, symmetry makes the expected posterior uniform. Increasing only the correct logit increases its posterior pointwise, while symmetry keeps all incorrect coordinate means equal. Hence the conditional mean has the form (142) with $\theta_v > 0$. Local JS reweighting is nonlinear and does not inherit zero-mean noise cancellation; it requires the additional control below.

The role of centering the collision matrix is also exact: an uncentered collision matrix adds the Gram matrix of view marginals, which can reward constant class bias rather than sample-dependent evidence.

Robust descent for the actual spectral–JS rule.

Proposition E.11 (Descent survives bounded spectral and posterior errors)**.** *Use the reference model above, but let $q_v(x)$ and $\widehat{C}$ be the actual emitted view posteriors and the collision matrix at the arrival prefix $n(x)$. Suppose the following bounds hold uniformly over all evaluated arrival prefixes:*

$$\begin{gathered}\operatorname*{ess\,sup}_{x,v} \text{TV}(q_v(x), q_v^0(x)) \leq \epsilon_q, \\ \widehat{C}_n = \sigma_{\text{sem}}^2\theta\theta^\top + \nu_n I + E_{C,n}, \qquad \sup_n \|E_{C,n}\|_{\text{op}} \leq \epsilon_C < g/2, \qquad g := \sigma_{\text{sem}}^2 S_2,\end{gathered} \tag{147}$$

where $\nu_n \geq 0$. Put

$$\begin{aligned}\delta_L &:= \frac{2\epsilon_C}{g - \epsilon_C}, & \epsilon_D &:= \sqrt{(\log 2)\epsilon_q}, \\ d_* &:= \sum_v L_v^* d(\theta_v) > 0, & h_* &:= \sqrt{\log 2}\,\delta_L + \epsilon_D.\end{aligned} \tag{148}$$

If $h_* < d_*$, the actual JS normalization is positive, and the three fusion stages satisfy $\text{TV}(p_a(x), p_a^0(x)) \leq \epsilon_a$ with

$$\epsilon_{\text{av}} = \epsilon_q, \qquad \epsilon_{\text{sp}} = \epsilon_q + \delta_L/2, \qquad \epsilon_{\text{full}} = \epsilon_q + h_*/d_*. \tag{149}$$

Consequently, on the same fixed graph,

$$e(p_b) - e(p_a) \leq -(b_b - b_a)\mathfrak{a}_{\text{sem}} + K_J(\epsilon_a + \epsilon_b). \tag{150}$$

Thus either fusion operation strictly decreases the actual realization gap whenever its ideal semantic gain exceeds the displayed perturbation cost.

Proof. At any evaluated prefix, abbreviate $\widehat{C}_n$, ν_n, and $E_{C,n}$ as $\widehat{C}$, ν, and E_C. Let u be its unit leading eigenvector and $u^* = \theta/\sqrt{S_2}$. Projecting its eigenvector equation onto the orthogonal complement of u^*, and using the Rayleigh lower bound $\lambda_1(\widehat{C}) \geq \nu + g - \epsilon_C$, gives $\sin\angle(u, u^*) \leq \epsilon_C/(g - \epsilon_C)$. Moreover, $\sum_v |u_v^2 - (u_v^*)^2| \leq \|u - u^*\|_2 \|u + u^*\|_2 = 2\sin\angle(u, u^*)$. Therefore $\|L - L^*\|_1 \leq \delta_L$.

The metric property of $\sqrt{\mathrm{JS}}$ (Endres & Schindelin, 2003) and $\mathrm{JS}(p, q) \leq (\log 2)\,\mathrm{TV}(p, q)$ imply $|D_v(x) - d(\theta_v)| \leq \epsilon_D$. The latter inequality follows by decomposing p, q into their common mass and disjoint residuals and applying joint convexity of JS; the residual JS is at most $\log 2$. Since $D_v \leq \sqrt{\log 2}$, the unnormalized evidence vectors satisfy $\|L \odot D - L^* \odot d(\theta)\|_1 \leq h_*$. Their reference sum is d_*, so the actual sum is positive and normalization gives $\|w - w^*\|_1 \leq 2h_*/d_*$.

A change of mixture weights contributes at most half their ℓ_1 distance to total variation; perturbing the mixture constituents contributes at most ϵ_q. This proves (149). Averaging within components preserves these bounds. Apply (135) to each actual–reference pair, then use (145). □

The collision errors $E_{C,n}$ may include anisotropic view noise, cross-view error correlations, sampling error, and drift among historical arrival-time posteriors. Their uniform bound is a condition, not an independence assertion about the deterministic views. Nonnegative semantic loadings are essential: a common wrong factor can have a large eigengap, and JS evidence can emphasize a confident error. Neither statistic alone supplies semantic orientation. Likewise, squared loading is shown to improve the stated calibrated objective in this regime, not to be a universally minimum-variance or Bayes-optimal weight.

Together, the results assign distinct roles to the modules: visibility removes irrelevant degrees of freedom exactly; centering removes identifiable offsets; transport reverses compatible anisotropy; orbit suppression trades lost signal against removed nuisance; and spectral–JS fusion increases correctly oriented shared evidence when its gain dominates perturbations. The condition $\mathfrak{a}_{\mathrm{sem}} > 0$ is an aggregate objective-descent condition and can hold without a positive worst-component identification margin. It does not extend the scope of Corollary E.2, which still requires $\Delta_{\mathrm{id}} > 0$. The separate latent-margin risk guarantee (138) requires its own semantic and restoration conditions. No claim of unconditional accuracy improvement is made from covariance normalization, agreement, or confidence alone.

E.5 Fixed-graph PAC–PRISM audit

Protocol. To avoid mixing the original ResNet-50 PAC audit with the ViT-B method evaluation, we build the bridge directly on the same frozen ViT-B source model used by PRISM. For each target domain, the component partition, masses, graph weights, source anchors, τ, λ, and kernel bandwidth are constructed once and then held fixed for all compared posteriors. The frozen configuration uses the identity source classifier-visible representation, $k = 20$, minimum component size 10, confidence core 0.4, a 128-dimensional Gaussian-RFF graph with bandwidth multiplier 0.1, $\tau = 0.7$, and $\lambda = 0.05$. We audit all 47 target domains: 15 CIFAR-10-C, 15 CIFAR-100-C, 15 ImageNet-C, ImageNet-R, and ImageNet-Sketch.

We use two sample scales. The short audit uses 256 samples for graph fitting and an independent 256-sample D_2 for posterior evaluation; the expanded audit uses 2048/2048. Target labels are revealed only after the graph and all predictions are frozen, to define $c(u)$, α_{tot}, Δ_{id}, M_{soft}, and accuracy. Labels never enter graph fitting, PRISM inference, or the LP defining $e(p)$. All LP optima are solver-certified by primal–dual bounds; the maximum interval width is 2.63×10^{-9} in the short audit and 5.55×10^{-9} in the expanded audit.

Source-to-PRISM bridge. Table 4 reports the corrected fixed-graph comparison. The accuracies are 47-domain audit-split averages and should not be confused with the five-benchmark macro averages in the main method table.

On the short audit, PRISM improves accuracy by 8.901 points and reduces the mean realization gap by 3.662×10^{-3}; the paired-bootstrap 95% interval for $e(\mathrm{PRISM}) - e(\mathrm{Source})$ is $[-4.560, -2.797] \times 10^{-3}$. The gap is lower in 42/47 domains, higher in one, and tied in four. On the expanded audit, accuracy improves by 8.949 points and the mean gap falls by 4.992×10^{-3}, with 95% interval

Table 4: Fixed-graph source-to-PRISM posterior calibration. Lower $\mathcal{J}_{\mathrm{TV}}$, $e(p)$, and M_{soft} are better.

Audit	Method	Acc. (%) ↑	$\mathcal{J}_{\mathrm{TV}}$ ↓	Mean $e(p)$ ↓	Median $e(p)$ ↓	M_{soft} ↓
512	Source	66.1735	0.00966563	0.00569256	0.00476141	0.183512
	PRISM	**75.0748**	**0.00600331**	**0.00203024**	**0.00145866**	**0.105601**
4096	Source	66.3553	0.0120834	0.00838086	0.00807933	0.282749
	PRISM	**75.3044**	**0.00709118**	**0.00338860**	**0.00286104**	**0.135051**

Table 5: Raw five-view averaging versus PRISM on the same frozen graphs. $\Delta e = e(\mathrm{PRISM}) - e(\mathrm{Raw5})$, so negative values indicate movement toward the calibrated semantic optimum.

Audit	$e(\mathrm{Raw5})$	$e(\mathrm{PRISM})$	Δe	Acc. gain
512	0.00479774	**0.00203024**	-0.00276750	$+5.718$
4096	0.00665267	**0.00338860**	-0.00326407	$+6.393$

$[-5.728, -4.263] \times 10^{-3}$; the gap is lower in $45/47$ domains and tied in the remaining two. Mean soft semantic mismatch simultaneously falls by 0.0779 and 0.1477, respectively.

Raw-five-view control. The same conclusion holds against naive equal-weight five-view averaging.

For the short audit, the paired-bootstrap 95% interval for Δe is $[-3.508, -2.045] \times 10^{-3}$; for the expanded audit it is $[-4.030, -2.509] \times 10^{-3}$. The realization gap is lower than raw-five-view averaging in $42/47$ and $45/47$ domains, respectively. Thus the bridge is not explained solely by replacing the single-view source with multiple views.

Risk-bound applicability. The risk corollary is used only on domains with positive empirical identification margin. Among the 17 such domains in the short audit, the unclipped right-hand side of (118) is below one in $16/17$ domains for PRISM, compared with $2/17$ for the source posterior. In the expanded audit, the corresponding counts are $12/15$ and $4/15$. These are finite-pool post-hoc audits, not deployment-time certificates or claims that every target domain satisfies the population assumptions.

Taken together, the PAC theory identifies a calibrated semantic optimum, the same objective accepts the actual PRISM posterior without modifying the algorithm, and PRISM moves substantially closer to that optimum while improving classification accuracy. This establishes a direct output-level bridge between recoverable semantic structure and the practical posterior emitted by PRISM, complementing the theory-guided design connection in Section 5.

F Detailed Related Work and Positioning

Expansion, self-training, and semantic propagation. Our recoverable-semantic-structure perspective is closely related to theoretical analyses of learning from unlabeled data under expansion or consistency conditions. Wei et al. (2020) analyze self-training with deep networks under an expansion assumption together with limited overlap across semantic classes. Cai et al. (2021) use a related expansion condition to establish label-propagation guarantees under subpopulation shift. These works establish that sufficiently connected same-class neighborhoods allow reliable supervision to propagate through unlabeled data. Our setting shares this propagation intuition but addresses an additional source-free TTA identification problem: the relevant component scale must itself be recoverable from target observations, source information must resolve the component-label permutation, and cross-semantic connections and wrong source evidence enter explicitly through the support–leakage margin Δ_{id}. The leave-one constructions in Appendix A.7 isolate these roles separately.

Self-training and entropy minimization under domain shift. Chen et al. (2020) prove that self-training and conditional-entropy minimization can avoid spurious source features under a structured domain-shift model, while Kumar et al. (2020) analyze self-training along gradually changing

unlabeled domains. These results provide important sufficient regimes for successful unsupervised adaptation. Our finite-width analysis asks a different mechanistic question: once a source model and target pool are fixed, where can a corrective direction arise when the entropy force on an isolated wrong prediction points in the wrong direction? The resulting answer is parameter sharing: source-oriented senders influence other samples through the finite-width Jacobian coupling kernel, yielding an explicit support-minus- leakage condition for early score correction.

Goyal et al. (2022) study TTA losses through convex conjugacy and show why entropy-like objectives arise naturally for cross-entropy-trained classifiers. This is complementary to our dynamics analysis. Their perspective concerns which unsupervised loss is compatible with the source training loss; ours decomposes how a fixed entropy objective can acquire a semantically correct update direction through interactions among target samples.

Graph-based test-time inference. LAME (Boudiaf et al., 2022) combines source predictions with an affinity-based Laplacian objective at test time and is therefore structurally related to the graph–anchor functional used in our PAC analysis. The role of our graph functional is different: it serves as an analyzable identification problem for characterizing when component semantics are recoverable, with explicit conditions for semantic-scale recovery, source orientation, leakage control, finite-sample exact identification, and leave-one obstructions. The practical method proposed in this paper is PRISM, rather than the graph-labeling solver used to establish the identifiability result.

Moment adaptation, recentering, and covariance alignment. Several adaptation methods correct distribution shift through low-order target statistics. Adaptive Batch Normalization (Li et al., 2016) replaces or modulates normalization statistics across domains, while Schneider et al. (2020) show that adapting batch-normalization statistics substantially improves robustness to common corruptions. CORAL and Deep CORAL (Sun & Saenko, 2016; Sun et al., 2017) align second-order feature statistics across domains. These methods motivate the general utility of statistical recentering and covariance correction, but PRISM uses these operations in a different information regime and geometry: correction is restricted to the classifier-visible subspace, all statistics are computed causally from the unlabeled target prefix without source-domain statistics, within-orbit covariance separates view-specific nuisance from shared variation, and the corrected views are subsequently fused through centered spectral consensus and local posterior evidence.

G Full Specification and Additional Evaluation of PRISM

This appendix gives a self-contained specification of PRISM and the experimental details omitted from Section 5. The method uses only a frozen classifier, deterministic target views, and causal statistics of the current target domain. It does not access source images, source feature statistics, target labels, future batches, or corruption identities, and it never updates model parameters.

G.1 Frozen classifier, fixed views, and causal target stream

Let the frozen network produce a penultimate feature $h \in \mathbb{R}^D$ and logits

$$s = Wh + b, \qquad W \in \mathbb{R}^{K\times D}, \quad b \in \mathbb{R}^K. \tag{151}$$

For a target domain, batches $\mathcal{B}_1, \mathcal{B}_2, \dots$ arrive sequentially. Prediction for batch t may use the current-batch-inclusive prefix

$$\mathcal{P}_t = \bigcup_{j=1}^{t} \mathcal{B}_j, \tag{152}$$

but no sample from any future batch. All target statistics are reset when the domain changes.

The standard ViT-B five-benchmark realization uses the same five deterministic views on every dataset: (i) a 90% center crop resized to the input resolution, (ii) Gaussian blur with radius 0.5, (iii) Gaussian blur with radius 1.0, (iv) a 3×3 median filter, and (v) an unsharp mask with radius 1, amount 150%, and threshold 2.

The construction follows the standard multi-view test-time augmentation principle used by methods such as MEMO and TPT: multiple label-preserving observations expose prediction invariances that

Table 6: Sensitivity to the deterministic view family under the same 512-sample evaluation protocol used in Table 1. Each variant changes exactly one view relative to the original five-view family, while the remainder of the adaptation pipeline is unchanged.

Dataset	Original	Variant A	Variant B	Variant C	Variant D
CIFAR-10-C	89.87	89.88	89.97	89.61	89.85
CIFAR-100-C	75.22	74.95	75.39	75.07	75.11
ImageNet-C	62.77	62.62	62.70	62.43	62.77
ImageNet-R	65.69	65.43	65.69	65.36	65.69
ImageNet-Sketch	53.26	53.39	53.26	53.26	53.13
Macro	69.36	69.25	69.40	69.15	69.31

are unavailable from a single view. Our operators deliberately span mild spatial, smoothing, and sharpening perturbations and are fixed globally rather than selected according to the corruption type. Unlike MEMO/TPT, however, PRISM does not optimize model parameters or prompts over these views; it treats them as repeated measurements for closed-form semantic consensus.

View-family variations. The original five-view family is `Crop90`, `GBlur0.5`, `GBlur1.0`, `Median3`, and `Unsharp`, in that order. Each variation changes exactly one view while keeping the other four fixed: Variant A replaces `Crop90` with `Crop95`; Variant B replaces `GBlur1.0` with `GBlur1.5`; Variant C replaces `Median3` with `BoxBlur1.0`; and Variant D replaces `Unsharp` with `Sharpness1.5`. Thus the five configurations are

	View 1	View 2	View 3	View 4	View 5
Original	Crop90	GBlur0.5	GBlur1.0	Median3	Unsharp
Variant A	Crop95	GBlur0.5	GBlur1.0	Median3	Unsharp
Variant B	Crop90	GBlur0.5	GBlur1.5	Median3	Unsharp
Variant C	Crop90	GBlur0.5	GBlur1.0	BoxBlur1.0	Unsharp
Variant D	Crop90	GBlur0.5	GBlur1.0	Median3	Sharpness1.5

The small variation across these view families indicates that the method does not rely on one privileged deterministic transform. We therefore keep the original five-view family fixed throughout the standard ViT-B five-benchmark evaluation. The separate cross-backbone transfer controls use the explicitly reported architecture-specific view budgets.

G.2 Classifier-visible coordinate system

Define the class-centered classifier

$$W_c = W - \frac{1}{K}\mathbf{1}\mathbf{1}^\top W. \tag{153}$$

Let its compact SVD be

$$W_c = U\Sigma B^\top, \tag{154}$$

where the columns of $B \in \mathbb{R}^{D\times r}$ correspond to nonzero singular values and $r = \operatorname{rank}(W_c) \le K-1$.

Proof of Proposition G.1. For $n \in \ker(W_c)$, $W_c n = 0$. Therefore the class-centered logits of $h+n$ and h coincide: $W_c(h+n) = W_c h$. Equivalently, Wn can differ from zero only by a class-shared constant, and softmax is invariant to adding the same scalar to all logits. Hence classifier-null feature directions cannot change the posterior, so every prediction-relevant correction may be represented in $\operatorname{span}(B)$. □

Define the classifier anchors

$$A = W_c B \in \mathbb{R}^{K\times r}. \tag{155}$$

For logits s_{iv} of sample i under view v, remove the classifier bias and class-shared logit component:

$$\widetilde{s}_{iv} = (s_{iv} - b) - \frac{1}{K}\mathbf{1}\mathbf{1}^\top(s_{iv} - b). \tag{156}$$

With $P = (A^\top)^\dagger$, the classifier-visible coordinate is

$$z_{iv} = \widetilde{s}_{iv} P \in \mathbb{R}^r. \tag{157}$$

The frame (B, A, P) is constructed from the frozen head alone. In the main ViT-B experiments, $r = 9$ for CIFAR-10-C, $r = 99$ for CIFAR-100-C, $r = 768$ for ImageNet-C and ImageNet-Sketch, and $r = 199$ for the 200-class ImageNet-R classifier mask.

G.3 Repeated-measurement model and causal centering

We use the repeated-measurement decomposition

$$z_{iv} = \mu_v + c_i + \varepsilon_{iv}, \tag{158}$$

where μ_v is a stable view offset, c_i is the component shared across views of sample i, and ε_{iv} is view-specific nuisance. For a visible prefix of n target images,

$$\mu_v = \frac{1}{n}\sum_{i=1}^{n} z_{iv}, \qquad \mu = \frac{1}{V}\sum_{v=1}^{V} \mu_v, \qquad \widehat{z}_{iv} = z_{iv} - \mu_v. \tag{159}$$

The view-specific means remove systematic offsets induced by the deterministic operators before any covariance correction is applied.

G.4 Total covariance and OAS stabilization

The prefix total covariance is

$$S_T = \frac{1}{nV}\sum_{i=1}^{n}\sum_{v=1}^{V} \widehat{z}_{iv}\widehat{z}_{iv}^\top. \tag{160}$$

The V views of one image are repeated measurements rather than independent target images, so the effective observation count in the OAS shrinkage rule is n, not nV.

Let

$$\bar{\lambda} = \frac{\mathrm{Tr}(S_T)}{r}, \qquad \alpha = \frac{\|S_T\|_F^2}{r^2}. \tag{161}$$

The shrinkage coefficient is

$$\rho = \min\left\{1, \frac{\alpha + \bar{\lambda}^2}{(n+1)(\alpha - \bar{\lambda}^2/r)}\right\}, \tag{162}$$

with $\rho = 1$ in the numerically degenerate denominator case. The stabilized covariance is

$$\Sigma_T = (1-\rho)S_T + \rho\bar{\lambda} I. \tag{163}$$

Let

$$\Sigma_T = U_T \,\mathrm{diag}(\lambda_1, \ldots, \lambda_r) U_T^\top, \qquad \lambda_{\mathrm{avg}} = \frac{1}{r}\sum_{j=1}^{r}\lambda_j. \tag{164}$$

Full isotropic matching would use $f_j = \sqrt{\lambda_{\mathrm{avg}}/\lambda_j}$. To avoid aggressive amplification of a low-variance direction when only a short prefix has been observed, define

$$\kappa_n = \frac{n}{n+r}, \qquad a_j = \begin{cases} f_j^{\kappa_n}, & \lambda_j < \lambda_{\mathrm{avg}}, \\ f_j, & \lambda_j \geq \lambda_{\mathrm{avg}}, \end{cases} \tag{165}$$

and

$$T_n = U_T \,\mathrm{diag}(a_1, \ldots, a_r) U_T^\top. \tag{166}$$

Thus high-variance directions are fully contracted toward the common scale, whereas low-variance directions are expanded gradually as the target prefix grows. Both ρ and κ_n are analytic functions of the unlabeled prefix; neither is selected per dataset.

G.5 WITHIN-ORBIT NUISANCE AND SYMMETRIC CONGRUENCE

For each image, define its view-average coordinate

$$\bar{z}_i = \frac{1}{V}\sum_{v=1}^{V} z_{iv}. \tag{167}$$

The across-image covariance of these shared coordinates is

$$S_B = \frac{1}{n}\sum_{i=1}^{n} \bar{z}_i \bar{z}_i^\top - \mu\mu^\top. \tag{168}$$

We estimate the within-orbit variation by

$$S_W = \mathrm{sym}(S_T - S_B). \tag{169}$$

Under (158), S_B is dominated by variation that persists across views of the same sample, whereas S_W isolates variation introduced within a view orbit.

Let

$$S_W = U_W \,\mathrm{diag}(\eta_1, \ldots, \eta_r) U_W^\top, \qquad \eta_{\mathrm{avg}} = \frac{1}{r}\sum_{j=1}^{r} \eta_j. \tag{170}$$

The nuisance multipliers are one-sided:

$$c_j = \min\left\{\sqrt{\frac{\eta_{\mathrm{avg}}}{\eta_j}}, 1\right\}, \qquad N_n = U_W \,\mathrm{diag}(c_1, \ldots, c_r) U_W^\top. \tag{171}$$

Hence directions with unusually large within-orbit variance are suppressed, while low within-orbit variance is never amplified. Standard numerical eigenvalue floors are used only to avoid division by zero and are fixed across all datasets.

The two corrections are combined by the symmetric congruence

$$G_n = N_n^{1/2} T_n N_n^{1/2}. \tag{172}$$

Let $t(i)$ denote the batch in which sample i arrives, and let $n_t = \sum_{s=1}^{t} |\mathcal{B}_s|$. For $i \in \mathcal{B}_t$, the current coordinate statistics define $\mu_v^{(t)}$ and $G_t := G_{n_t}$. The posterior emitted for this sample is

$$\widetilde{q}_{iv} := \mathrm{softmax}\Big((z_{iv} - \mu_v^{(t)}) G_t A^\top + b\Big). \tag{173}$$

Equivalently, $\widetilde{q}_{iv}$ is evaluated using the transform available at the arrival prefix $n_{t(i)}$. Once emitted, this posterior remains fixed as later batches modify the coordinate statistics. At this point no pseudo-label has been constructed; the remainder of the method uses the full emitted posterior vectors.

G.6 CENTERED COLLISION COVARIANCE AND SPECTRAL LEVERAGE

The posterior statistics are accumulated over the emitted posterior stream, not over a hypothetical re-evaluation of the full prefix under the latest transform. Let

$$Q_v^{(t)} = Q_v^{(t-1)} + \sum_{i \in \mathcal{B}_t} \widetilde{q}_{iv}, \qquad R_{vu}^{(t)} = R_{vu}^{(t-1)} + \sum_{i \in \mathcal{B}_t} \widetilde{q}_{iv}^\top \widetilde{q}_{iu}, \tag{174}$$

with $Q_v^{(0)} = 0$ and $R_{vu}^{(0)} = 0$. The causal posterior marginal and centered collision covariance after batch t are therefore

$$m_v^{(t)} = \frac{Q_v^{(t)}}{n_t}, \qquad C_{vu}^{(t)} = \frac{R_{vu}^{(t)}}{n_t} - (m_v^{(t)})^\top m_u^{(t)}. \tag{175}$$

To see why this quantity measures shared view structure, suppose

$$\widetilde{q}_{iv} - m_v = a_v g_i + \xi_{iv}, \qquad g_i \in \mathbb{R}^K, \tag{176}$$

where g_i is a centered semantic factor shared by all views and the cross-view centered noises satisfy $\mathbb{E}[\xi_{iv}^\top \xi_{iu}] = 0$ for $v \neq u$ and are uncorrelated with g_i. Then, for $v \neq u$,

$$\mathbb{E}[(\widetilde{q}_{iv} - m_v)^\top (q_{iu} - m_u)] = a_v a_u \, \mathbb{E}[\|g_i\|_2^2]. \tag{177}$$

Thus the population collision matrix contains a rank-one shared component proportional to $aa^\top$ (up to view-specific diagonal noise). When this component has a leading eigengap, the leading eigenvector identifies the loading direction up to its global sign, and u_v^2 is the corresponding variance-share score. We use this interpretation only to motivate the closed-form weighting; we do not claim Bayes optimality of the finite-sample fusion rule. The cross-view-independent noise condition is a sufficient interpretation of the rank-one factor model, not an empirical assumption imposed on the fixed crop, blur, median-filter, and sharpening views. These deterministic views can exhibit correlated errors. We therefore do not treat the leading spectral direction as a certificate of semantic correctness; its utility is evaluated empirically through the ablations below.

G.7 Per-sample Jensen–Shannon evidence and closed-form fusion

Let $u_K = (1/K, \ldots, 1/K)$ and define

$$\mathrm{JS}(\widetilde{q}_{iv}, u_K) = \frac{1}{2}\,\mathrm{KL}\left(\widetilde{q}_{iv} \middle\| \frac{\widetilde{q}_{iv} + u_K}{2}\right) + \frac{1}{2}\,\mathrm{KL}\left(u_K \middle\| \frac{\widetilde{q}_{iv} + u_K}{2}\right). \tag{178}$$

Because the square root of Jensen–Shannon divergence is a metric on probability distributions (Lin, 1991; Endres & Schindelin, 2003), define the local evidence amplitude

$$D_{iv} = \sqrt{\mathrm{JS}(\widetilde{q}_{iv}, u_K)}. \tag{179}$$

The evidence score, view weight, fused posterior, and prediction are

$$e_{iv} = L_v D_{iv}, \qquad w_{iv} = \frac{e_{iv}}{\sum_{j=1}^{V} e_{ij}}, \qquad p_i = \sum_{v=1}^{V} w_{iv} \widetilde{q}_{iv}, \qquad \widehat{y}_i = \arg\max_k p_{ik}. \tag{180}$$

This is the complete decision rule. There is no learned gate, expert router, confidence threshold, rescue stage, or gradient update.

G.8 Causal algorithm and information boundary

For each target domain, PRISM executes:

1. reset all target-prefix statistics;
2. for the incoming batch $\mathcal{B}_t$, compute logits for the fixed deterministic views with the frozen model;
3. construct (B, A, P) from the frozen classifier head and map the current view logits to classifier-visible coordinates z_{iv};
4. include the current-batch coordinates in the prefix coordinate sufficient statistics, giving $\mu_v^{(t)}, S_T^{(t)}, S_B^{(t)}, S_W^{(t)}$;
5. compute OAS stabilization, T_t, N_t, and $G_t = N_t^{1/2} T_t N_t^{1/2}$;
6. compute the current-batch arrival-time posteriors $\widetilde{q}_{iv}$ using Equation (173);
7. add only these newly emitted posteriors to $Q_v^{(t)}$ and $R_{vu}^{(t)}$ in Equation (174); historical posterior contributions remain fixed;
8. form $C^{(t)}$, extract $L_v = (u_v^{(t)})^2$, compute $D_{iv} = \sqrt{\mathrm{JS}(\widetilde{q}_{iv}, u_K)}$, and fuse the current batch using $w_{iv} \propto L_v D_{iv}$;
9. proceed to the next batch. Future batches never revise an earlier prediction or posterior contribution.

Thus the coordinate statistics and posterior statistics have different semantics. The former define the transform available at the current prefix, whereas the latter summarize the sequence of posteriors actually emitted by the online algorithm. In particular, the implementation never attempts to reconstruct historical softmax outputs under an updated transform from first- or second-order coordinate moments.

The method reads only frozen model parameters and current/past unlabeled target batches. It does not read source images, source means, source covariance matrices, source prototypes, target labels, future batches, or corruption names. The principal online state consists of second moments in the r-dimensional visible space, view-coordinate sums, posterior marginal sums, and a $V \times V$ collision matrix, giving

$$O(r^2 + Vr + VK + V^2) \tag{181}$$

memory. This is the state required by the online method itself. The diagnostic implementation additionally archives emitted posteriors for later analysis, which costs $O(nVK)$ but is not required for inference. The fixed classifier-visible maps require an additional $O(Kr)$ storage. The dominant post-processing operation is a constant number of symmetric $r \times r$ eigendecompositions per update,

$$O(r^3). \tag{182}$$

The $O(r^3)$ term describes the dominant matrix-decomposition component of the post-processing cost, rather than the complete end-to-end inference cost.

G.9 Experimental protocol and baseline provenance

The main ViT-B experiments use batch size 64. For CIFAR-10-C, CIFAR-100-C, and ImageNet-C, we evaluate severity 5 under the standard 15 corruptions using 512 target samples per corruption. ImageNet-R and ImageNet-Sketch use the corresponding 512-sample target evaluation with 3 seeds. Every corruption or natural shift resets the online state.

The main ViT-B benchmark comparison in Table 1 reports the mean over three fixed evaluation seeds. The corresponding standard deviations are 0.13, 0.62, 0.48, 0.18, and 0.79 percentage points on CIFAR-10-C, CIFAR-100-C, ImageNet-C, ImageNet-R, and ImageNet-Sketch, respectively.

Table 1 defines the standard complete PRISM configuration under the 512-sample ViT-B protocol. The complete-method entries in the view-sensitivity, component-ablation, and supplementary benchmark tables refer to this same evaluation configuration. Numerical differences correspond to explicit changes in the evaluated configuration, such as removing a component, modifying the deterministic view family, or applying a controlled intervention.

Information boundary. Target labels are used only for final performance evaluation. The deterministic view family, closed-form correction rule, spectral aggregation rule, and all inference-time method settings are fixed without using target labels. The adaptation procedure itself accesses only the frozen model and causal unlabeled target statistics.

For ImageNet-C, CIFAR-10-C, ImageNet-R, and ImageNet-Sketch, the baseline entries in Table 1 use the published ViT-B results of Murphy et al. (2026). CIFAR-100-C is evaluated under the corresponding standardized protocol. For controlled comparisons, we also evaluate the baselines under a common implementation. For NEO, we report additional matched-sample, long-stream, and matched-view comparisons in Appendix G.10.

G.10 Matched-Protocol Baseline Evaluation and View Controls

The experiments in this subsection distinguish three comparison settings. First, Table 1 uses the published baseline values from Murphy et al. (2026) for direct literature comparison. Second, we evaluate the established baselines under our standardized evaluation pipeline; NEO is also evaluated under the matched protocol reported below. Third, because the standard ViT-B realization of PRISM uses five deterministic views, we construct a separate NEO-5V control using exactly the same five views. The standard NEO and NEO-5V settings are distinct controls and are reported separately.

Standardized evaluation of non-NEO baselines. Table 7 reports the remaining baselines under our standardized ViT-B evaluation pipeline. We report NEO separately below because it is additionally evaluated under a stricter sample-level paired protocol against PRISM.

On the four benchmarks shared with the published comparison (ImageNet-C, CIFAR-10-C, ImageNet-R, and ImageNet-Sketch), the standardized evaluation results are close to the literature values and are generally slightly more conservative. Averaged over these four benchmarks, they are lower than the corresponding published values by 1.07, 0.61, 1.14, 1.13, 0.44, 1.32, 0.66, and 0.84 percentage

Table 7: Established baselines under our standardized ViT-B test-time adaptation protocol. Published reference values used for the literature comparison are reported separately in Table 1.

Dataset	No Adapt	T3A	SAR	LAME	Tent	CoTTA	FOA	Surgeon
ImageNet-C	54.82	54.05	54.67	53.11	56.88	54.71	57.22	54.93
CIFAR-10-C	79.21	80.72	79.48	78.62	80.55	79.22	79.81	81.45
ImageNet-R	58.01	57.55	58.40	57.36	58.59	57.88	60.94	58.85
ImageNet-Sketch	44.27	44.73	44.21	43.88	44.92	43.82	45.18	47.40
CIFAR-100-C	61.56	63.03	61.56	55.51	61.39	61.60	66.80	58.88
Avg	59.57	60.02	59.66	57.70	60.47	59.44	61.99	60.30

Table 8: Matched-protocol comparison of NEO and PRISM under the same 512-sample ViT-B evaluation setting as Table 1.

Dataset	NEO (matched)	PRISM	Gain
CIFAR-10-C	84.57	**89.87**	+5.30
CIFAR-100-C	67.53	**75.22**	+7.68
ImageNet-C	58.07	**62.77**	+4.70
ImageNet-R	59.24	**65.69**	+6.45
ImageNet-Sketch	45.70	**53.26**	+7.56
Macro	63.03	**69.36**	**+6.33**

points for No Adapt, T3A, SAR, LAME, Tent, CoTTA, FOA, and Surgeon, respectively. CIFAR-100-C is evaluated under the same standardized pipeline and has no corresponding published entry in this comparison.

Matched-protocol evaluation of NEO. Because NEO is the strongest and most directly related optimization-free baseline, we evaluate it under the same 512-sample ViT-B protocol used for PRISM. The PRISM column reports the three-seed mean under this standard configuration. This table is a matched-protocol comparison rather than a per-seed paired-difference estimate.

For the four datasets that also have published NEO results in Table 1, the published versus matched-protocol accuracies are 82.40 versus 84.57 on CIFAR-10-C, 59.20 versus 58.07 on ImageNet-C, 60.30 versus 59.24 on ImageNet-R, and 47.20 versus 45.70 on ImageNet-Sketch. Their four-dataset macro averages are 62.28% and 61.90%, respectively. The matched-protocol values are slightly lower on three of the four overlapping datasets while remaining close to the published reference overall.

Long-stream evaluation. We next test whether the improvement is specific to the initial 512-sample stream. Using seed 2020, we evaluate 4,096 non-repeated target samples per domain; the first 512 samples preserve the corresponding short-stream indices and order.

The gain therefore persists beyond the initial 512 samples: after excluding them entirely, the macro improvement over NEO remains +4.80 percentage points. This experiment uses one fixed seed and is not interpreted as a multi-seed long-stream estimate.

Strict five-view control. Because the standard ViT-B realization of PRISM uses five deterministic views, we additionally construct a five-view extension of NEO to control for the number of backbone forwards. NEO-5V applies the standard NEO correction independently to the same five deterministic view streams used by PRISM and fuses the resulting posteriors with equal weights. Thus, this comparison holds the view family and forward-view budget fixed while changing only the adaptation and fusion rule.

The advantage therefore remains at the aggregate level when the number and identity of forward views are explicitly matched. The short-stream PRISM value is the three-seed mean under the standard 512-sample protocol. NEO-5V denotes the five-view control obtained by applying NEO to the same deterministic view family used by PRISM.

Table 9: Long-stream matched-protocol evaluation with $4{,}096$ samples per target domain. The final column excludes the first 512 samples.

Dataset	Source	NEO (matched)	PRISM	Full gain	Gain on 513–4096
CIFAR-10-C	79.32	85.20	**89.85**	+4.65	+4.55
CIFAR-100-C	61.32	67.17	**74.56**	+7.39	+7.37
ImageNet-C	55.06	58.63	**63.48**	+4.85	+4.89
ImageNet-R	60.60	61.23	**64.50**	+3.27	+2.82
ImageNet-Sketch	44.95	46.90	**51.64**	+4.74	+4.35
Macro	6.25	63.83	**68.81**	**+4.98**	**+4.80**

Table 10: Strict five-view comparison between NEO-5V and PRISM. Both methods use the same five deterministic views.

Evaluation window	NEO-5V	PRISM-5V	Gain
Short stream, 512 samples	63.96	**69.36**	**+5.40**
Long stream, 4096 samples	64.03	**68.81**	**+4.78**

Cross-backbone equal-view controls. Finally, we perform additional equal-view comparisons on other architectures. These controls use the same number of forward views within each matched-view comparison and are separate from the published architecture baselines in Table 2.

Together, these experiments provide both a literature-reference comparison and controlled evaluations under a common implementation. The improvement persists under the matched protocol, over the extended target stream, and when the number of forward views is explicitly matched.

G.11 Component ablations

The decomposition separates two effects that are conflated by a prediction-side fusion ablation. Multi-view inference supplies additional observations, whereas the subsequent stages progressively remove systematic view shift and nuisance before estimating shared semantic consensus. In particular, the classifier-visible projection itself does not create a prediction gain; its role is to remove classifier-null degrees of freedom from the statistics used by the later corrections.

G.12 Full ImageNet-C comparison by corruption

Table 12 augments the published ImageNet-C comparison with PRISM. For compactness, we list the published means and omit the confidence intervals from the original table.

PRISM exceeds the published NEO mean on $13/15$ corruption types. The two exceptions are fog and JPEG compression; this prevents the aggregate gain from being attributable to a uniformly favorable rescaling of every corruption.

G.13 Hard online robustness on ImageNet-C

We additionally stress-test the frozen PRISM implementation on ImageNet-C under more difficult online streams without modifying its views, covariance estimator, nuisance correction, spectral aggregation, or inference rule. All experiments use the same severity-5 benchmark and the same archived 512-sample pool per corruption as the standard evaluation. Before running the stress tests, the standard episodic $B = 64$ prediction archive is reproduced exactly; its 15-corruption average is 62.773%, matching the seed-specific corruption results in Table 12 and rounding to the same 62.77% reported in the standard summary tables.

To keep numerical provenance consistent across the paper, the robustness table below reports PRISM accuracies and within-PRISM changes only. The main comparison tables retain their separately specified baseline provenance, so no baseline value from a different evaluation pipeline is recombined here.

Table 11: Stepwise decomposition of PRISM under the standard 512-sample evaluation protocol. All multi-view variants use exactly the same five target views. The complete PRISM row corresponds to the standard complete configuration in Table 1. The classifier-visible re-expression is prediction invariant by Proposition G.1; preceding rows isolate the contribution of view centering, total-covariance transport, within-orbit nuisance suppression, global spectral leverage, and local JS evidence.

Variant	C10-C	C100-C	IN-C	IN-R	IN-S	Macro
Frozen source	79.21	61.56	54.82	58.01	44.27	59.57
Raw five-view average	81.25	66.41	56.05	60.16	48.04	62.38
+ Classifier-visible re-expression	81.25	66.41	56.05	60.16	48.04	62.38
+ Total-covariance transport	86.18	71.46	60.66	62.70	50.39	66.28
+ View-wise centering	88.32	73.22	61.65	64.65	50.98	67.76
+ Within-orbit suppression	89.54	74.39	62.73	64.45	51.95	68.61
+ Centered spectral leverage	89.02	74.96	62.34	65.43	52.54	68.86
+ Local JS evidence (PRISM)	**89.87**	**75.22**	**62.77**	**65.69**	**53.26**	**69.36**

Table 12: ImageNet-C severity-5 accuracy (%) by corruption with ViT-B. The first nine columns are the published means from Murphy et al. (2026); the last column is PRISM with seed 2020.

Corruption	No Adapt	T3A	SAR	LAME	Tent	CoTTA	FOA	Surgeon	NEO	PRISM
Gaussian noise	57.0	56.7	57.0	56.5	57.2	57.0	57.2	58.7	57.7	**58.79**
Shot noise	56.9	57.0	57.3	56.8	57.5	57.1	58.6	58.8	57.6	**60.94**
Impulse noise	57.4	57.0	57.5	56.7	57.6	57.7	57.7	58.9	58.1	**65.23**
Defocus blur	46.9	47.5	47.5	47.1	48.0	48.2	49.2	49.1	49.8	**56.05**
Glass blur	35.3	35.9	36.3	34.9	36.8	36.0	36.8	36.8	37.9	**39.84**
Motion blur	53.3	53.1	53.5	52.9	54.0	53.3	54.4	54.8	55.0	**58.98**
Zoom blur	44.8	45.3	46.3	44.7	46.4	45.0	47.1	45.7	47.5	**54.69**
Snow	62.2	62.7	62.6	58.5	63.0	63.2	64.3	62.2	64.6	**67.38**
Frost	62.6	63.3	63.3	62.2	63.3	63.0	63.9	61.7	65.0	**70.31**
Fog	65.8	62.9	65.4	62.0	62.4	64.8	70.7	63.0	**71.2**	69.92
Brightness	77.9	78.1	78.0	78.1	78.2	77.9	78.2	78.1	78.3	**78.91**
Contrast	32.6	27.5	34.0	24.9	36.9	33.2	54.5	31.7	58.2	**60.94**
Elastic transform	45.8	45.8	45.7	44.4	46.7	46.3	49.6	46.2	49.8	**57.23**
Pixelate	67.5	67.4	67.5	67.1	68.0	67.8	67.2	67.6	68.2	**74.61**
JPEG compression	67.9	67.3	67.3	66.9	67.6	67.8	68.6	68.9	**69.1**	67.77
Average	55.6	55.2	55.9	54.2	56.3	55.9	58.5	56.1	59.2	**62.77**

The small-batch result is stable across the full stream: reducing the batch size from 64 to 1 changes average accuracy by only 0.61 percentage points, and batch sizes 2, 4, and 8 remain within 0.49 points of the standard setting. Under ordering perturbations, the class-blocked and sticky streams reduce accuracy by 0.98 and 0.82 points relative to the same-sample IID $B = 8$ control. Finally, retaining the online state across all 15 corruption segments reduces accuracy from 62.77% episodically to $61.79\% \pm 0.11$ over three fixed corruption orders. Thus the frozen method retains most of its standard ImageNet-C accuracy under the tested small-batch, ordering, and no-reset continual settings, without any stress-specific adaptation rule or retuning.

G.14 From the Support–Leakage Principle to PRISM

PRISM is not a solver for the discrete PAC objective and is not claimed to optimize Δ_{id} or Δ_J. The direct theory–method interface is instead the posterior calibration of Appendix E, which applies the PAC semantic structure to the actual probabilistic output of PRISM and yields a conditional risk bound through $e(p)$. The present subsection explains the separate, method-specific roles of classifier visibility and repeated-view statistics.

Shared variation across deterministic views is not assumed to be semantically correct: multiple views may share the same error, while view-specific variation may still contain useful discriminative information. The repeated-measurement model below therefore explains the spectral estimator, not the semantic correctness of its leading factor."

Table 13: ImageNet-C hard online robustness for the frozen PRISM implementation. Accuracy is averaged over the 15 severity-5 corruptions. For ordering experiments, Δ is measured against the same-sample IID control at the same batch size. Continual accuracy is mean$\pm$sample standard deviation over three fixed corruption orders. The Dirichlet row changes both the realized class prior and the available sample count and is therefore reported without a standard-protocol delta.

Setting	PRISM Acc.	Reference	Δ (pp)
Standard episodic, $B = 64$	62.77	standard protocol	0.00
Episodic, $B = 1$	62.16	$B = 64$	-0.61
Episodic, $B = 2$	62.28	$B = 64$	-0.49
Episodic, $B = 4$	62.36	$B = 64$	-0.42
Episodic, $B = 8$	62.34	$B = 64$	-0.43
Dirichlet 0.1, $B = 8$	62.81	different stream	—
IID shuffle, $B = 8$	62.93	IID control	0.00
Class-blocked, $B = 8$	61.95	IID $B = 8$	-0.98
Sticky 0.95, $B = 8$	62.11	IID $B = 8$	-0.82
Continual, no reset, $B = 64$	61.79 ± 0.11	episodic $B = 64$	-0.99

The correspondence has two formal components. First, prediction-relevant correction can be restricted exactly to the classifier-visible subspace.

Proposition G.1 (Classifier-visible sufficiency). *Let $W_c = W - K^{-1}\mathbf{1}\mathbf{1}^\top W$. For every $n \in \ker(W_c)$ and feature h,*

$$\mathrm{softmax}(W(h+n)+b) = \mathrm{softmax}(Wh+b).$$

Hence every posterior-relevant feature correction has an equivalent representation in the row space of W_c.

Proof. $W_c n = 0$ implies that Wn is constant across classes. Adding a common scalar to all logits leaves softmax unchanged. □

Second, repeated measurements provide an observable decomposition between shared evidence and view-specific variation. Consider

$$\widetilde{q}_{iv} - m_v = a_v g_i + \xi_{iv}, \qquad \mathbb{E}[g_i] = 0, \tag{183}$$

where g_i is shared across views, a_v is the loading of view v, and ξ_{iv} is centered view-specific noise. Assume $\mathbb{E}[g_i^\top \xi_{iv}] = 0$ and $\mathbb{E}[\xi_{iv}^\top \xi_{iu}] = 0$ for $v \neq u$. Then for $v \neq u$,

$$\mathbb{E}\big[(\widetilde{q}_{iv} - m_v)^\top (q_{iu} - m_u)\big] = a_v a_u \,\mathbb{E}\,\|g_i\|_2^2 . \tag{184}$$

Thus the population collision matrix contains a rank-one shared component proportional to $aa^\top$, together with view-specific diagonal noise. If the shared component induces a leading eigengap, its principal eigenvector is aligned with the shared loading direction; isotropic diagonal noise gives exact recovery up to sign, while bounded anisotropic diagonal noise produces the standard eigengap-controlled perturbation.

This result explains the role of spectral leverage without identifying the finite-sample leading factor with ground-truth semantics. Semantic correctness is supplied instead by the calibrated PAC interface in Appendix E, where the actual emitted posterior is evaluated against the source-oriented semantic optimum.

G.15 Five-benchmark results and corruption-level breakdown

G.16 Ablations and theory-aligned diagnostics

Under this sequential decomposition, spectral leverage contributes approximately $+0.25$ percentage points beyond equal fusion, and the complete local-JS formulation reaches 69.36%, corresponding to a further $+0.50$ points beyond spectral leverage. These are sequential conditional increments and are not interpreted as order-independent additive effects.

Table 14: PRISM accuracy on the five primary benchmarks under the standard 512-sample ViT-B protocol, corresponding to the complete configuration in Table 1.

Dataset	Accuracy (%)
CIFAR-10-C	89.87
CIFAR-100-C	75.22
ImageNet-C	62.77
ImageNet-R	65.69
ImageNet-Sketch	53.26
Macro average	**69.36**

Table 15: Prediction-side fusion ablation under the same standard 512-sample evaluation protocol as the stepwise decomposition. All rows use the same frozen checkpoint, ordered target samples, five deterministic views, batch size, and arrival-time statistic updates. The complete PRISM row uses the standard complete configuration from Table 1.

Variant	Macro accuracy (%)
Corrected five-view stream + equal fusion	68.61
+ centered spectral leverage	68.86
+ local JS evidence (PRISM)	**69.36**

The leave-one-view-out result is below the full five-view construction, indicating that the gain does not depend on a single privileged transform.

These statistics quantify posterior concentration and cross-view agreement only. A wrong prediction may also become more confident and thereby increase the top-1–top-2 margin or its distance from the uniform posterior. Consequently, we do not interpret these quantities as semantic certificates or identification-margin lower bounds. Semantic alignment is evaluated instead by the calibrated fixed-graph audit in Appendix E.

The nuisance intervention is a leave-one-module-out test of the otherwise complete pipeline and should therefore not be numerically identified with the 68.61% sequential-ablation row, which has not yet added the later spectral and local-JS stages.

G.17 Computational Efficiency

The matched-protocol baseline evaluations and matched-view accuracy controls are reported together in Appendix G.10. Here we isolate the computational cost of those evaluations. Although PRISM requires no backward graph, parameter gradients, or optimizer state, its multi-view inference and spectral post-processing incur non-negligible additional computation.

The raw measurements make clear that the principal practical cost of PRISM is multi-view backbone inference plus spectral post-processing. Accordingly, "optimization-free" denotes the absence of test-time backpropagation and parameter optimization, rather than zero computational overhead. Matched-view accuracy controls are reported separately in Appendix G.10.

Table 16: Posterior-concentration diagnostics before and after PRISM. The last column is the Spearman correlation between diagnostic change and accuracy gain across target domains. These statistics are descriptive and are not lower bounds on Δ_{id} or Δ_J.

Diagnostic	Positive domains	Mean change	Spearman
Mean top-1–top-2 margin	47/47	+0.03282	0.2146
JS-distance lower-tail	47/47	+0.02445	0.3683

Table 17: Separate counterfactual interventions on the full PRISM pipeline. Unlike the sequential ablation in Table 15, the "none" nuisance-suppression row removes that module while retaining the subsequent spectral and local-JS fusion stages; the two tables therefore represent different configurations.

Intervention	Strength	Macro accuracy (%)
Within-orbit nuisance suppression	none	69.091
	natural	**69.36**
	$4\times$	69.003
Cross-sample semantic mixing	0%	**69.36**
	25%	67.50–67.60

Table 18: Raw end-to-end evaluation cost per target domain. "Alloc./res." denotes peak allocated/reserved GPU memory.

Backbone	Method	Views	Wall time (s)	Alloc./res. (MiB)	Post-process (s)
ResNet-50	NEO-1V	1	1.605	279/342	0.029
ResNet-50	PRISM-5V	5	8.779	812/1046	3.603
ViT-S	NEO-4V	4	4.231	375/520	0.013
ViT-S	PRISM-4V	4	5.439	382/544	1.143
ViT-L/16	NEO-4V	4	23.379	1858/2222	0.011
ViT-L/16	PRISM-4V	4	25.971	1881/2282	1.787